Weighted Association Rules and Scoring Methodology for Cardiovascular Diseases

Gokhan Goy[1*], Burak Kolukisa[1], Burcu Bakir-Gungor[1], Ibrahim Ugur[2], Vehbi Cagri Gungor[1]

[1] Deparment of Computer Engineering, Abdullah Gül University, Sümer Campus, Kayseri 38080, Turkey.
[2] Keydata Bilgi İşlem Teknoloji Sistemleri A.Ş., Ankara, Turkey.

* Corresponding author. Tel: +90 553 669 38 46; email: gokhan.goy@agu.edu.tr

Manuscript submitted January 25, 2019; accepted May 28, 2019.
doi: 10.17706/ijbbb.2019.9.4.222-230

Abstract: Cardiovascular diseases (CVD), including coronary artery disease (CAD), myocardial infarction, and stroke are a group of highly prevalent and deadly diseases. The deaths from cardiovascular diseases were announced as 17.9 million in 2016 and it is expected that this number will reach approximately to 23.6 million by 2030. In order to facilitate the diagnosis and treatment of CVD, several computational approaches and data mining methods have been proposed until now. In this study, Apriori algorithm is utilized to find associations between features and rules based on UCI's publicly available Cleveland dataset. Additionally, we generate different weighted association rules, which can help medical doctors to stratify patients and thus, propose different treatment approaches for each patient's sub-category. Performance results show that the Apriori algorithm creates 58 rules when support and confidence parameters are set to 0.1 and 0.9, respectively. Utilizing weighted association rule approach, 6 important rules have been created based on Clinical Important Factors (CIF) and Framingham Heart Study Risk factors (FHS RF) on CVD.

Key words: Association rule, cardiovascular diseases, rule scoring, weighted association rules.

1. Introduction

Cardiovascular diseases (CVD) involving multiple risk factors, pathological changes in diverse cell types, tissues, and organs, and multidimensional molecular perturbations cause increasing numbers of deaths worldwide. According to World Health Organization (WHO), 31% of the world's total deaths in 2016 (17.9 million) was due to CVD [1]. It is expected that the number of deaths due to CVD, will increase to approximately 30 million by 2030. For this reason, this topic has attracted the attention of a large number of researchers and several studies have been conducted in this area. In general, these studies can be classified into the following six sub-categories: *Diagnosis, Prognosis, Screening, Monitoring, Management, Treatment.*

- *Diagnosis*: Karabulut & İbrikçi have provided a system that helps decision-making with the Rotation Forest algorithm and diagnoses coronary artery disease [2].
- *Prognosis*: De Falco proposed a methodology based on differential evolution for classifying items automatically [3].
- *Screening*: Jen *et al*. aimed to establish an early warning system in order to reduce deaths resulted from chronic diseases [4].
- *Monitoring*: Shen and Kao proposed a system for cardiac arrhythmia detection by using feature selection and support vector machines in electrocardiograms (ECG) [5].

- *Management*: Kusiak *et al*. developed a metric for analyzing data and assessing quality measurement and its benefits [6].
- *Treatment*: Idri *et al*. attempted to diagnose a disease that was specifically identified in their study and to recommend appropriate treatment [7].

Researchers studied different classifiers and feature selection methods to diagnose and prognose CVDs. In Table, these studies have been compared and summarized. Different from these existing studies, our focus in this study is to find relationships between features and their impacts on CVDs. To address this need, Apriori algorithm is utilized to find associations between features and rules based on UCI's publicly available *Cleveland* dataset [8]. In this study, we also generate different weighted association rules, which can help medical doctors to stratify patients and thus propose different treatment approaches for each patient sub-category. The Apriori algorithm is a globally accepted algorithm that finds frequent items and then generates robust association rules from these items. However, the Apriori algorithm tends to give an overwhelming number of rules. To this end, the interestingness measure, which indicates which rule is more important and more difficult to detect, is used to filter out unimportant rules [9]. Also, the Apriori algorithm assumes that all features have an equal impact on diseases, but in reality this is not the case for CVDs. For example, the impacts of verified risk factors, such as the factors announced by Framingham Heart Study (FHS) [10] or Clinically Important Factors (CIF), are more significant for CVDs compared to other features, as shown in Table 2. Framingham Heart Study is conducted by Boston University and National Heart, Lung and Blood Institute (NHBLI) [10]. To address this issue, we used weighted association rule mining and considered both computational values and weights given by FHS. This paper is organized as follows. In section 2, the evaluated approaches and development environment have been explained. In Section 3, performance results have been shown. Finally, the paper is concluded in Section 4.

Table 1. A Summary of Data Mining Approaches to CVD

Study	Algorithm	Category	Dataset	Year
Karabulut & İbrikçi [2]	Rotation Forest	Diagnosis	Cleveland	2012
R. Das *et al.* [11]	Ensemble	Diagnosis	Cleveland	2009
De Falco [3]	Differential Evolution	Prognosis	UCI	2013
Hsieh *et al.* [12]	Ensemble	Prognosis	Private	2012
Jen *et al.* [4]	kNN, Linear Discriminant Analysis	Screening	Private	2012
Podgorelec *et al.* [13]	Ensemble	Screening	Private	2005
Shen *et al.* [5]	SVM	Monitoring	MIT-BIH arrhythmia	2012
Czabanski *et al.* [14]	Ensemble	Monitoring	Private	2012
Kusiak *et al.* [6]	Rough Set Theory	Management	Private	2006
Antonelli *et al.* [15]	DBSCAN	Management	Private	2013
Idri *et al.* [7]	KNN, C4.5, Random Forest, Naïve Bayes, SVM	Treatment	Private	2017

Table 2. Feature Types

Feature No	Feature Type	Feature
1	CIF	Cp
2	CIF	Exang
3	CIF	Oldpeak
4	CIF	Thal
5	FHS RF	Trestbps
6	FHS RF	Chol
7	FHS RF	Fbs
8	FHS RF	Age
9	FHS RF	Gender

2. Materials and Methods

In this section, the algorithms and the development environments used in this study have been explained.

2.1. Association Rule Mining

Association Rule Mining is a method that seeks to uncover hidden and potentially valuable association rules from the data set. The terminology of association rule mining is as follows: Let $F = \{f_1, f_2, \ldots, f_{n-1}, f_n\}$ be a set of features (features that have categorical values), and $\overline{T} = \{t_1, t_2, \ldots, t_{k-1}, t_k\}$ be a set of transactions (all records). Transactions dataset is described by $\overline{T}$, where each $t_k \in \overline{T}$ contains a set of features $F' \subseteq F$. In the light of this statement, an association rule can be expressed as ***'antecedent (A) → consequent (C)'*** where $A, C \subset F$ and $A \cap C = \emptyset$. More specifically, the Apriori algorithm uses two important parameters when it finds the frequent items. These parameters are support and confidence parameters. These two parameters are required to generate rules that can be called robust or confident. The standardized formulas of these parameters are shown below in as follows.

$$Support(x) = \frac{Frequency(x)}{Number\ of\ All\ Transactions} \tag{1}$$

$$Confidence(x \Longrightarrow y) = \frac{Support(xy)}{Support(x)} = \frac{Frequency(xy)}{Frequency(x)} \tag{2}$$

2.2. Apriori Algorithm

Apriori algorithm is one of the most widely used algorithms of association rule mining. The Apriori algorithm is a globally accepted algorithm that finds frequent items and then generates robust association rules from these items [16]. If a feature or feature group has a bigger support value than the support threshold, then the feature or feature group is counted frequently; and if the rules created with these frequent features exceed the given confidence threshold, this rule is reported as important and acceptable.

2.3. Development Environment

One of the development environments that we use in this study is the *R* development environment [17]. *R* is a free development environment that includes data analysis and statistical computing. The operations in *R* are performed through packages. The packages used in this work are *arules* [18] and *NbClust* [19] packages. arules package is an R package developed to determine the frequent elements and to construct their association rules. NbClust package determines how many pieces of this data group should be best separated. In order to do this, it uses various values previously mentioned in the literature. Another development environment that we used in this study is WEKA [20]. WEKA platform includes several machine learning algorithms and with the help of its user-friendly interface, it is widely used in data mining tasks. It can be used from its own interface or used directly in JAVA programming language if desired.

2.4. Dataset

The dataset used in this research effort is the Cleveland dataset from UCI machine learning repository. Feature descriptions of Cleveland dataset are shown in Table 3.

2.5. Proposed Approach

The Apriori algorithm assumes that all features have an equal impact on the outcome, but in reality, this is not the case for CVDs. For example, the impacts of Clinically Important Factors (CIF) and Framingham Heart Study Risk Factors (FHS RF) on CVD are more significant compared to other features. To address this issue, we used weighted association rule mining and considered both computational values and weights given by FHS. In order to assign weights to association rules, firstly, the weight of each item needs to be found. To calculate these weights, both the computational value and the weights given by FHS were taken into account. The flowchart of our approach is shown in Fig. 1.

Table 3. Feature Descriptions

No	Feature	Description
1	Gender	Relevant Person's gender (male or female)
2	Fbs	Fasting Blood Sugar >120 mg/dl (0 or 1)
3	Exang	Exercise Induced Angina (Yes or No)
4	Cholesterol	Serum Cholesterol in mg/dl
5	Age	Relevant Person's age in years
6	Trestbps	Resting Blood Pressure
7	Restecg	Resting Electrocardiographic Result
8	Thalach	Maximum Heart Rate Achieved
9	Slope	The Slope of The Peak Exercise ST Segment
10	Thal	Normal, Fixed Defect, Reversable Defect
11	Cp	Chest Pain Type
12	Oldpeak	St Depression Induced By Exercise Relative To Test
13	Ca	Number Of Major Vessels Colored By Flourosopy
Class	Class	Diagnosis of Heart Disease

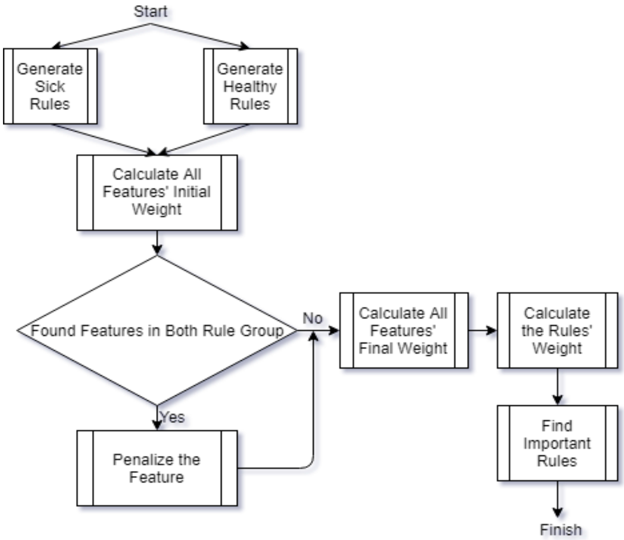
Start
Generate Sick Rules
Generate Healthy Rules
Calculate All Features' Initial Weight
Found Features in Both Rule Group
No
Yes
Penalize the Feature
Calculate All Features' Final Weight
Calculate the Rules' Weight
Find Important Rules
Finish

Fig. 1. Flow chart of the proposed approach.

In Fig. 1, while the *sick rules* indicate the information of the disease in the consequent part of the rule, *healthy rules* indicate that there is no disease information in the consequent part of the rule. In our proposed approach, firstly, the sick and healthy rule groups were generated using the arules package. Secondly, the initial weights of each feature were calculated via checking the frequency of a feature among all generated rules, as shown in as follows.

$$Initial\ Weight\ of\ a\ Feature = \frac{Frequency\ in\ Rules(Feature)}{Number\ of\ Rules} \tag{3}$$

Thirdly, if a feature exists in both healthy and sick rule groups, to increase that feature's effect on the disease, the two values are substracted from each other, and hence, feature is penalized. Fourthly, the final weight of the feature is calculated by processing the weights according to FHS. This process is done according to as follows.

$$Final\ Weight\ of\ a\ Feature = \frac{Initial\ Weight * Given\ Weight}{2} \tag{4}$$

Finally, the individual weight of each rule is calculated. This weight also determines the score of the relevant rule. This is done according to as follows.

$$Weight\ of\ A\ Rule = \frac{Sum\ of\ Final\ Weights\ of\ Feature}{Number\ of\ Features} \tag{5}$$

After this process, we determine the most important rules according to the scores and finish the algorithm.

3. Performance Results

In this study, UCI's publicly available *Cleveland* dataset is utilized, since it is the most comprehensive dataset and includes a limited number of missing values compared to other datasets. *Cleveland* dataset contains 303 samples and 14 features.

3.1. Data Preparation

In *Cleveland* dataset, real numbers need to be converted into categorical values to be able to use in the Apriori algorithm. To achieve this, we set the data in the coordinate system using all the properties of WEKA's visual abilities, and we determined the points of breakage, where the data are going to change. These values are shown in Table 4.

Table 4. Categorization Values of Attributes

Feature	Features' Range	Nominal New Value
Gender	Male, Female	1, 2
Fbs	True, False	1, 2
Exang	Yes, No	1, 2
Cholesterol	X < 265	1, 2
Age	X < [55,66] < Y	1, 2, 3
Trestbps	X < [145, 163] < Y	1, 2, 3
Thalach	X < [98, 112] < Y	1, 2, 3
Slope	Up, Flat, Down	1, 2, 3
Ca	1, 2, 3, 4	1, 2, 3, 4
Restecg	Normal ST_T_wave_abnormality Left_Vent_Hyper	1, 2, 3
Thal	Normal, Fixed Defect Reversable Defect	3, 6, 7
Cp	Typical Angina Atypical Angina Non Anginal Pain Aysmtomatic	1, 2, 3, 4
Oldpeak	X< 1 1 < Y < 2 2 < Z < 3 T > 3	1, 2, 3, 4

3.2. Association Rule Mining Results & Filtering Process

The most important weakness of the Apriori algorithm is that it generates an excessive number of rules due to the fact that it operates according to the frequency of the data. Here, the most important problem is to understand which of the generated rules contain more important hidden data. The support, confidence and score of a rule are used to address this problem. We created the rules via applying the Apriori algorithm using *arules* package. In this study, we have used the support parameter as 0.1 in order not to miss the important rules when generating the association rules. We have generated all rules by increasing the confidence parameter by 0.1 starting from 0.5 to 0.9. However, we have noticed that most of these rules didn't have a medical meaning, so we only used the rules generated when the confidence parameter was 0.9. In cases where the support value for the sick and healthy rules is 0.1 and the confidence value is 0.9, the number of generated rules is 58 and 3302, respectively.

3.3. Feature Weights

In our proposed approach, the weights of each feature were calculated separately. This was done by taking into account both FHS considerations and feature ranking results information. Here, we only calculated the weights of feature values observed in sick rules. The initial and final weights are given in Table 5.

Table 5. Features' Weight

Feature	Initial Weight	Final Weight
Gender = 1	0,1565	0,21
Fbs = 2	-0,0548	-0,07
Exang = 1	0,2423	0,345
Cholesterol = 2	0,0234	0,03
Age = 2	0,1667	0,23
Trestbps = 1	-0,12	-0,15
Restecg = 3	0,1176	0,14
Thalach = 3	-0,1164	-0,13
Slope = 2	0,2635	0,31
Thal = 7	0,3238	0,48
Cp = 4	0,4218	0,63
Oldpeak = 2	0,0183	0,033
Ca = 2	0,0305	0,034

When the scores of the rules generated with these weights were analyzed, it was observed that these values varied between 0.1 and 0.5. One of the values used in the selection of rules in this study was the score of a rule. The rules with a score bigger than 0.4 were found to be meaningful by the medical doctors. These rules are given in Table 6.

Table 6. Meaningful Weighted Rules

Rule Number	Antecedent	Consequent	Support	Confidence	Score
3	{Cp=4 ∩ Thalach=3 ∩ Slope=2 ∩ Thal=7}	Sick	0.1353	0.9761	0.42
9	{Cp=4 ∩ Exang=1 ∩ Slope=2 ∩ Thal=7}	Sick	0.1056	0.9696	0.44
16	{Cp=4 ∩ Slope=2 ∩ Thal=7}	Sick	0.1518	0.9583	0.47
19	{Cp=4 ∩ Gender=1 ∩ Slope=2 ∩ Thal=7}	Sick	0.1188	0.9473	0.41
28	{Cp=4 ∩ Gender=1 ∩ Exang=1 ∩ Thal=7}	Sick	0.1353	0.9318	0.42
45	{Cp=4 ∩ Age=2 ∩ Thal=7}	Sick	0.1353	0.9111	0.44
51	{Cp=4 ∩ Exang=1 ∩ Slope=2}	Sick	0.1617	0.9074	0.43

4. Conclusion

Cardiovascular diseases (CVD) are one of the leading causes of death in the world. Since CVDs account for 31% of world's annual deaths [1], they demand a better understanding. Until now, researchers studied different classifiers and feature selection methods to improve classification accuracy. Our focus in this study is explore the relationships between features and association rules and the quality of these generated rules with their own scores. To address this need, in this study, we first examined the relationships between the association rules for CVDs. To this end, we created association rules from UCI's publicly available Cleveland dataset using the Apriori algorithm. While generating these rules, we set support value as 0.1 and confidence value as 0.9. Using these two parameters we aimed to eliminate unnecessary and medically implausible rules. In addition to all these computational processes, each feature was given a separate weight in order to improve its medical meaning. Based on medical doctor's recommendations, we conclude that the rules with a score higher than 0.4 are reliable in medical applications. Future work includes the investigation of the proposed approach with different datasets.

Acknowledgment

This work was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) TEYDEB progam under Project no 3180177.

References

[1] Cardiovascular diseases (CVDs). *World Health Organization*. Retrieved from the website: http://www.who.int/mediacentre/factsheets/fs317/en/

[2] Karabulut, E. M., & İbrikçi, T. (2012). Effective diagnosis of coronary artery disease using the rotation forest ensemble method. *Journal of Medical Systems*, *36(5)*, 3011-3018.

[3] Falco, I. (2013). Differential evolution for automatic rule extraction from medical databases. *Applied Soft Computing*, *13(2)*, 1265-1283.

[4] Jen, C. H., Wang, C. C., Jiang, B. C., Chu, Y. H., & Chen, M. S. (2012). Application of classification techniques on development an early-warning system for chronic illnesses. *Expert Systems with Applications*, *39(10)*, 8852-8858.

[5] Shen, C. P., Kao, W. C., Yang, Y. Y., Hsu, M. C., Wu, Y. T., & Lai, F. (2012). Detection of cardiac arrhythmia in electrocardiograms using adaptive feature extraction and modified support vector machines. *Expert Systems with Applications*, *39(9)*, 7845-7852.

[6] Kusiak, A., Caldarone, C. A., Kelleher, M. D., Lamb, F. S., Persoon, T. J., & Burns, A. (2006). Hypoplastic left heart syndrome: Knowledge discovery with a data mining approach. *Computers in Biology and Medicine*, *36(1)*, 21-40.

[7] Idri, A., & Kadi, I. (2017, August). A data mining-based approach for cardiovascular dysautonomias diagnosis and treatment. *Proceedings of 2017 IEEE International Conference on Computer and Information Technology (CIT)* (pp. 245-252).

[8] UCI 2018 Heart Disease dataset. (2018). Retrieved from the website: https://archive.ics.uci.edu.ml/ heart+Disease

[9] Tan, P. N., Kumar, V., & Srivastava, J. (2002, July). Selecting the right interestingness measure for association patterns. *Proceedings of the Eighth ACM SIGKDD İnternational Conference on Knowledge Discovery and Data Mining* (pp. 32-41). ACM.

[10] Framingham: Past and present. (2018). Retrieved from the website: https://www.framinghamheartstudy.org/

[11] Das, R., Turkoglu, I., & Sengur, A. (2009). Effective diagnosis of heart disease through neural networks ensembles. *Expert Systems with Applications*, *36(4)*, 7675-7680.

[12] Hsieh, N. C., Hung, L. P., Shih, C. C., Keh, H. C., & Chan, C. H. (2012). Intelligent postoperative morbidity prediction of heart disease using artificial intelligence techniques. *Journal of Medical Systems*, *36(3)*, 1809-1820.

[13] Podgorelec, V., Kokol, P., Stiglic, M. M., Heričko, M., & Rozman, I. (2005). Knowledge discovery with classification rules in a cardiovascular dataset. *Computer Methods and Programs in Biomedicine*, *80*, S39-S49.

[14] Czabanski, R., Jezewski, J., Matonia, A., & Jezewski, M. (2012). Computerized analysis of fetal heart rate

signals as the predictor of neonatal acidemia. *Expert Systems with Applications*, *39(15)*, 11846-11860.

[15] Antonelli, D., Baralis, E., Bruno, G., Cerquitelli, T., Chiusano, S., & Mahoto, N. (2013). Analysis of diabetic patients through their examination history. *Expert Systems with Applications*, *40(11)*, 4672-4678.

[16] Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. *Acm Sigmod Record*, *22(2)*, 207-216.

[17] Team, R. C. (2013). R: A language and environment for statistical computing.

[18] Hahsler, M., Chelluboina, S., Hornik, K., & Buchta, C. (2011). The arules R-package ecosystem: Analyzing interesting patterns from large transaction data sets. *Journal of Machine Learning Research*, *12(Jun)*, 2021-2025.

[19] Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A., & Charrad, M. M. (2014). Package 'nbclust'. *Journal of Statistical Software*, *61*, 1-36.

[20] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, *11(1)*, 10-18.

Gokhan Goy received his B.S. degree in computer engineering from Erciyes University, Kayseri, Turkey, in 2017. Currently, he is a PhD student and research assistant, Abdullah Gül University (AGU), Kayseri, Turkey. His current research interests are in data mining, machine learning and bioinformatics.

Burak Kolukisa received his B.S. degree in computer engineering from Erciyes University, Kayseri, Turkey, in 2016. Currently, he is a MSc student and research assistant, Abdullah Gül University (AGU), Kayseri, Turkey. His current research interests are in data mining, machine learning, digital image processing and computer vision.

Burcu Bakir-Gungor received her B.Sc. degree in biological sciences and bioengineering from Sabanci University; her M.Sc. degree in bioinformatics from Georgia Institute of Technology; and her PhD degree from Georgia Institute of Technology/Sabanci University. Now, she works as an assistant professor in the Department of Computer Engineering at Abdullah Gül University. Her research interests include bioinformatics, computational genomics, network and pathway oriented analysis of genome-wide and applications of machine learning.

İbrahim Uğur received the B.Sc degrees in electrical and electronics engineering from Kırıkkale University, Kırıkkale, Turkey in 2010. He received the M.S degree in management ınformation systems (MIS) from Gazi University, Ankara, Turkey in 2017. He worked as a software specialist and software consultant in different companies for 8 years He has been working as technology director in Keydata ınformation technologies for 2 years. He has microsoft MCSD software specialist certificate.

Vehbi Cagri Gungor received his B.S. and M.S. degrees in electrical and electronics engineering from Middle East Technical University, Ankara, Turkey, in 2001 and 2003, respectively. He received his Ph.D. degree in electrical and computer engineering from Georgia Institute of Technology, Atlanta, GA, USA, in 2007. Currently, he is a full professor and chair of Computer Engineering Department, Abdullah Gül University (AGU), Kayseri, Turkey. His current research interests are in smart grid communications, machine-to-machine communications, next-generation wireless networks, wireless sensor networks.