Cardiac Studies Diagnostic Data Informative Features Investigation based on Cumulative Frequency Analysis

Kseniia Bazilevych[1], Mykola Butkevych[1] and Nataliia Dotsenko[2]

[1] *National Aerospace University "Kharkiv Aviation Institute", Chkalow str., 17, Kharkiv, 61070, Ukraine*
[2] *O.M. Beketov National University of Urban Economy in Kharkiv, Marshal Bazhanov str., 17, Kharkiv, 61002, Ukraine*

Abstract

The development of information technology in the modern world affects the public health sector on the one hand and accumulates enormous amounts of data on the other hand. The global COVID-19 pandemic has contributed to the digitalization of healthcare. Heart disease is a global problem that causes death worldwide. Therefore, this study proposes a model for determining the information content of signs of diagnostic data of heart diseases based on the cumulative frequency method. The software implementation of the model has been completed. A database of 303 patients, consisting of 14 attributes, was used for the experiments. As a result of the model's work, the features with the most significant information content were identified. The study is promising and can apply diagnostic models in public health practice.

Keywords 1

Features informativeness, cumulative frequency analysis, medical diagnostics, heart disease, data-driven medicine

1. Introduction

Cardiovascular disease is the leading cause of adult death worldwide. Mortality reaches 30% of the total number of all deaths [1]. Cardiovascular diseases are congenital and acquired. The following are distinguished among cardiovascular diseases [2]:

- Arterial hypertension.
- Cardiac ischemia.
- Acute coronary syndrome.
- Heart disease.
- Heart failure.
- Arrhythmia.
- Venous thrombosis.
- Atherosclerosis.

The main danger of cardiovascular disease is the disability or sudden death. The likelihood of such consequences increases when ignoring the signs of the disease. Among the main risk factors are [3]:

- Smoking.
- Alcohol abuse.
- Lack of physical activity.
- Unbalanced nutrition.
- Stress.

2nd International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2022), December 2-4, 2022, Łódź, Poland
EMAIL: ksenia.bazilevich@gmail.com (K. Bazilevych); nikolai.butkevych@gmail.com (M. Butkevych); nvdotsenko@gmail.com (N. Dotsenko)
ORCID: 0000-0001-5332-9545 (K. Bazilevych); 0000-0001-8189-631x (M. Butkevych); 0000-0003-3570-5900 (N. Dotsenko)

Also, the causes of cardiovascular diseases include high blood pressure and diabetes. Therefore, early diagnosis is one of the most effective methods of preventing cardiovascular diseases.

The COVID-19 pandemic has stimulated research in the field of data-driven medicine to solve various problems. These areas include modeling the epidemic process of infectious diseases [4, 5], the study of molecular structures [6], the study of social factors affecting the spread of disease [7], the study of the behavior of viruses [8], medical diagnostics [9], etc.

However, the available data on the disease does not always allow the construction of high-quality models of automated medical diagnostics.

This study aims to determine the information content of signs for the diagnosis of cardiovascular diseases using the cumulative frequency method.

Given research is part of a complex intelligent information system for epidemiological diagnostics, the concept of which is discussed in [10].

2. Materials and Methods

2.1. Features informativeness

Often the data sets to be processed contain a large number of features. When building machine learning models, it is not always clear which of the features are important for it and which are redundant [11]. At the same time, the removal of redundant data allows a better understanding of the data, as well as reducing the time for setting up the model, improving its accuracy and facilitating interpretability. Often this is the most important task. Feature selection methods are divided into three types:

- Filter methods.
- Embedded methods.
- Wrapped methods.

The choice of the appropriate method is not obvious and depends on the data.

In the field of data-driven medicine, it is possible to recognize the presence or absence of a disease only when certain signs inherent in the patient are received and analyzed. Such signs are called informative [12]. But informative features are not equivalent to achieve a specific goal, so determining their informativeness is an important task.

Informativeness of a sign means how much this sign characterizes the state of the object, that is, how much the diagnosis depends on it - the result of recognition. At the same time, two approaches can be distinguished for determining the information content: energy and information.

The energy approach is based on the fact that the information content is estimated by the value of the feature. However, this approach may be poorly suited for object recognition. If some attribute is large in absolute value, but almost the same for objects of different classes, then it is difficult to attribute the object to a certain class by the value of this attribute. And if the attribute is relatively small in size, but differs greatly for objects of different classes, then the object can be easily classified by its value.

According to the informational approach, feature information is considered as a reliable difference between classes of images in the feature space. When classifying objects, such a significant difference can be the difference in the probability distributions of a feature built on samples from comparable classes.

2.2. Cumulative frequency method

The essence of the cumulative frequency method is that if there are two samples of a feature x belonging to two different classes, then for both samples in the same coordinate axes, there are empirical distributions of the feature x [13]. The cumulative frequencies are calculated, i.e. the sum of frequencies from the initial to the current distribution interval. In this case, the module of the maximum difference of the accumulated frequencies serves as an estimate of information content:

$$I(x) = \max_{j=0,..,q} \left| M_{1j} - M_{2j} \right|, \tag{1}$$

where M_{1j} is the cumulative frequency for the j-th sampling interval A_1;
M_{2j} is the cumulative frequency for the j-th sampling interval A_2;
q + 1 is the number of intervals.
The cumulative frequency algorithm is shown in Figure 1.

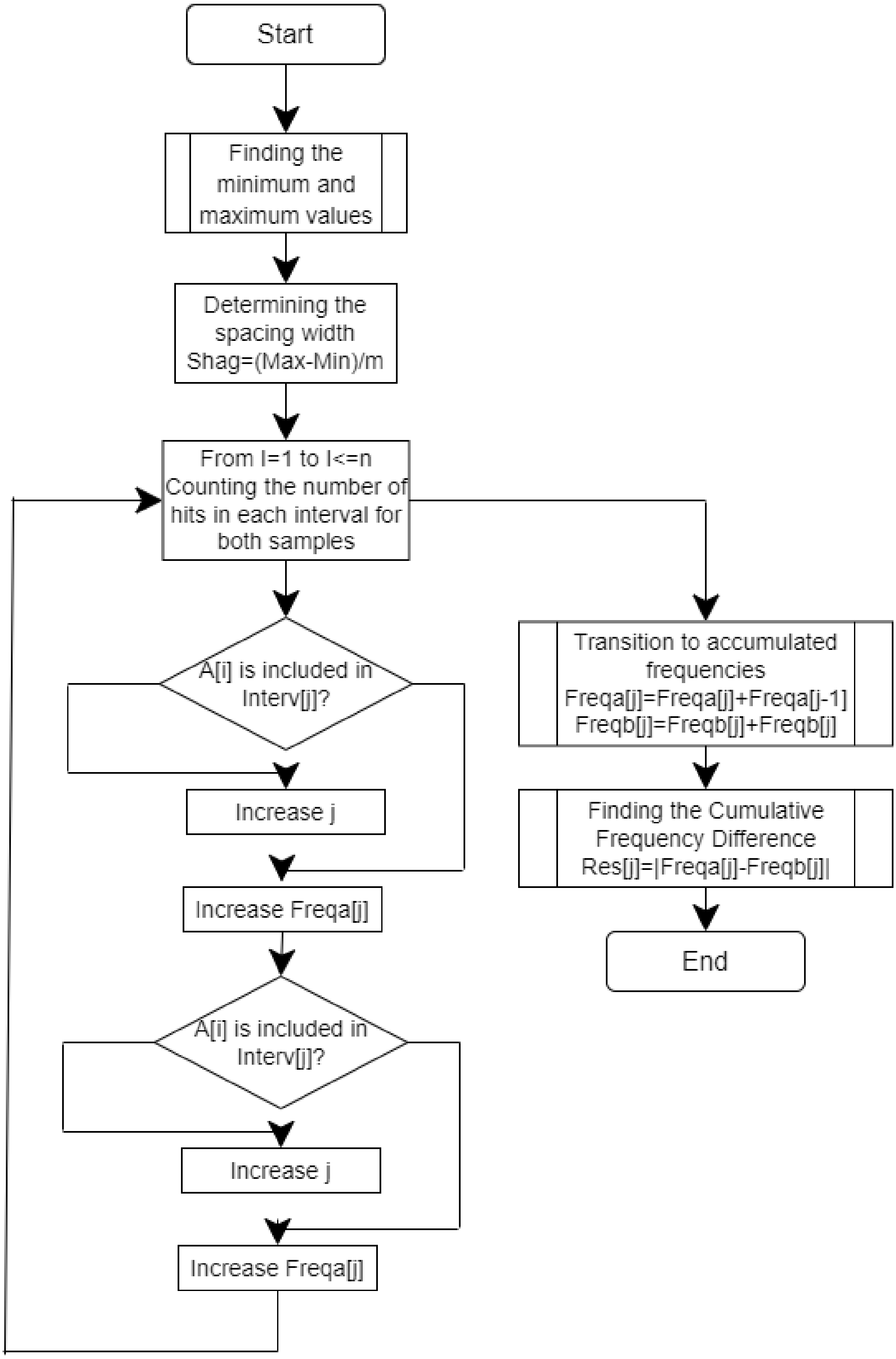

Figure 1: The algorithm of the cumulative frequency method.

3. Results

Experimental studies were carried out using the Python programming language. The open Heart Disease Cleveland dataset [14] was used for the analysis. The dataset contains data on 303 patients with 14 attributes. Attribute data is shown in Table 1.

Table 1
Description of the data

Attribute	Description
Age	Age in years
Sex	Sex (1=male; 0=female)
Chest pain type	1: typical angina; 2: atypical angina; 3: non-anginal plan; 4: asymptomatic
Blood pressure	Resting blood pressure
Cholesterol	Serum cholesterol in mg/dl
Fasting blood sugar < 120	1=trye; 2=false
Resting ECG	0: normal; 1: having ST-T wave abnormality; 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
Maximum heart rate	Maximum heart rate achieved
Angina	Exercise included angina (1=yes; 0=no)
Peak	ST depression induced by exercise relative to rest
Slope	The slope of the peak exercise ST segment
Colored vessels	Number of major vessels (0-3) colored by flourosopy
Thal	3=normal; 6=fixed defect; 7=reversable defect
Predicted attribute	0: <50% diameter narrowing; 1: >50% diameter narrowing

Data was distributed to two classes: "Healthy" and "Sick".

The results of informative features by cumulative features method are presented in Table 2.

As a result, the information content was calculated for different groups of cardiological data. It was found that the following signs are the most informative: thal, chest pain type, colored vessels, angina, age. The cumulative frequency method is used to determine the information content of a feature involved in the recognition of two classes of objects.

The use of an automated software package developed in the framework of this study allows its use at workplaces in medical institutions to support decision-making when making a diagnosis. An automated solution is especially relevant in conditions of limited resources in low- and middle-income countries and during force majeure, such as war, natural disasters, and other conditions in which access to medical care is limited.

Table 2
Description of the data

Attribute	Result
Age	19
Sex	-1
Chest pain type	50
Blood pressure	37
Cholesterol	6
Fasting blood sugar < 120	45
Resting ECG	147
Maximum heart rate	11
Angina	99
Peak	21
Slope	142
Colored vessels	66
Thal	118

4. Conclusions

As a result of the study, an automated software package was used to determine the information content of the signs of these patients with suspected heart disease based on the accumulated frequency method. An open dataset of patients with suspected heart disease was used for experimental studies, which included 303 patients and 14 attributes. It was found that the following signs are the most informative: thal, chest pain type, colored vessels, angina, and age.

The proposed software package is highly relevant in Russia's war in Ukraine, as it does not require high computing power. At the same time, automating a doctor's diagnosis and decision-making support in conditions of limited resources is an urgent task.

5. Acknowledgements

The study was funded by the National Research Foundation of Ukraine in the framework of the research project 2020.02/0404 on the topic “Development of intelligent technologies for assessing the epidemic situation to support decision-making within the population biosafety management”.

6. References

[1] M. Alessandro, P.E. Puddu, Epidemiology of heart disease of uncertain etiology: a population study and review of the problem, Medicina 55 (10) (2019): 687. doi: 10.3390/medicina55100687

[2] S.M. Hollenberg, Valvular heart disease in adult: etiologies, classification, and diagnosis, FP essentials 457 (2017): 11-16.

[3] S.S. Virani, et. al., Heart disease and stroke statistics – 2020 update: a report from the American heart association, Circulation 141 (9) (2020): e139-e596. doi: 10.1161/CIR.0000000000000757

[4] D. Chumachenko, et. al., On agent-based approach to influenza and acute respiratory virus infection simulation, 14th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering, TCSET 2018 – Proceedings (2018): 192-195. doi: 10.1109/TCSET.2018.8336184

[5] D. Chumachenko, On intelligent multiagent approach to viral hepatitis B epidemic processes simulation, Proceedings of the 2018 IEEE 2nd International Conference on Data Stream Mining and Processing, DSMP 2018 (2018): 415-419. doi: 10.1109/DSMP.2018.8478602

[6] A.S. Tkachenko, et. al., Semi-refined carrageenan promotes generation of reactive oxygen species in leukocytes of rats upon oral exposure but not in vitro, Wiener Medizinische Wochenschrift 171 (3-4) (2021): 68-78. doi: 10.1007/s10354-020-00786-7

[7] N. Davidich, et. al., Monitoring of urban freight flows distribution considering the human factor, Sustainable Cities and Society 75 (2021): 103168. doi: 10.1016/j.scs.2021.103168

[8] D. Chumachenko, K. Chumachenko, S. Yakovlev, Intelligent simulation of network work propagation using the Code Red as an example, Telecommunications and Radio Engineering 78 (5) (2019): 443-464. doi: 10.1615/TELECOMRADENG.V78.I5.60

[9] I. Izonin, R. Tkachenko, I. Dronyuk, R. Tkachenko, M. Gregus, M. Rashkevych, Predictive modeling based on small data in clinical medicine: RBF-based additive input-doubling method, Mathematical Biosciences and Engineering 18 (3) (2021): 2599-2613. doi: 10.3934/mbe.2021132.

[10] S. Yakovlev, et. al., The concept of development a decision support system for the epidemic morbidity control, CEUR Workshops Proceedings 2753 (2020): 265-274.

[11] W. Jitkrittum, et. al., Informative features for model comparison, Advances in Neural Information Processing Systems 31 (2018): 1-12.

[12] T. Tran, et. al. A framework for feature extraction from hospital medical data with applications in risk prediction, BMC Bioinformatics 15 (2014): 425. doi: 10.1186/s12859-014-0425-8

[13] M. Riachi, J. Himms-Hagen, M.E. Harper, Percent relative cumulative frequency analysis in indirect calorimetry: application to studies of transgenic mice, Canadian journal of physiology and pharmacology 82 (12) (2004): 1075-83. doi: 10.1139/y04-117

[14] R. Detrano, et. al., International application of a new probability algorithm for the diagnosis of coronary artery disease, American Journal of Cardiology 64 (1989): 304-310.