DOI: https://doi.org/10.31341/jios.43.1.1 UDC 004.855.5:005.53
Open Access Original Scientific Paper

Discretization for Naive Bayes Taking the Specifics of Heart Data into Account

Jan Bohacik *jan.bohacik@fri.uniza.sk*
Faculty of Management Science and Informatics
University of Zilina, Zilina, Slovakia

Michal Zabovsky *michal.zabovsky@fri.uniza.sk*
University Science Park
University of Zilina, Zilina, Slovakia

Abstract

At the present time heart disease is a major cause of death. Factors such as physical inactiveness, obesity, diabetes, social isolation and aging are expected to make the situation worse. It is worsened even further with misdiagnosis of patients describing heart related issues. A probability decision support approach to diagnosis of heart disease based on Naive Bayes is discussed here as most hospitals collect patient records but these are rarely used for automatic decision support. The approach is analyzed on Statlog heart data with the focus on improving preprocessing methods. As the result, a discretization algorithm with Equal Frequency Discretization which considers the specifics of engaged heart disease patients is presented. Enhancements of achieved accuracy with the added discretization and in comparison with other machine learning algorithms are shown in experiments founded on 10-fold cross-validation.

Keywords: discretization, Naive Bayes, diagnosis, heart disease

1. Introduction

The term "heart disease" is related to several disorders which are associated with the heart [11]. The most common of them is coronary artery disease and it is often connected with heart attack. Other typical disorders are difficulties of the heart to work as the pump, heart arrhythmia, rheumatic heart disease or valves in the heart. Heart disease stays with the patient once it appears and thus it is a lifelong condition. Some people are even born with it. Known risk factors for heart disease include diabetes, high blood pressure, high cholesterol, obesity, physical inactiveness, smoking, and social isolation [10]. The term is often used interchangeably with the term "cardiovascular disease" which includes conditions related to blocked or narrowed blood vessels as well. About 17.5 million people die from cardiovascular disease each year based on the data from World Health Organization [21], which is approximately one third of all deaths worldwide. More than two thirds of the deaths happen in lowincome and middle-income countries [20]. This is because the poor have difficulties to access or afford preventive services and ongoing treatments [3]. The ratio of risk factors such as obesity, physical inactiveness and smoking is also rising in low and middle-income countries [2]. The signs and symptoms of heart disease depend on the particular disorder involving the heart [11]. Emergency signs include chest discomfort, fainting and shortness of breath [18]. Other symptoms are: a) anxiety; b) discomfort radiating to the arm, back, jaw, or throat; c) indigestion; e) nausea and vomiting; and f) sweating. However, there are many situations with no symptoms at all. For instance, almost half of all heart attacks may cause no obvious symptoms [23]. But this silence does not mean that these heart attacks are less dangerous. Although silent heart attacks may accidently be found sometimes when the doctor is testing for something else, this is not sure at all and it is likely to be too late. It is therefore necessary to detect the occurrence of heart disease for high risk people as accurately as possible and as quickly as possible and give them effective help immediately.

Decisions are made on the basis of the experience traditionally. The doctor looks for signs and symptoms and makes use of screening tests. Typical tests include measurements of the amount of cholesterol, electrocardiography, exercise cardiac stress test, measurements of the blood glucose level and blood pressure level [15]. Combining found signs, symptoms and results of the tests is not trivial and poor clinical decisions based on the experience may lead to unwanted biases and medical errors [14]. These biases and errors could be reduced by integration of clinical decision support with computer-based patient records which helps to diagnose heart disease and avoid misdiagnosis [22]. Most hospitals collect patient records in information systems nowadays but these are rarely used for automatic intelligent decision support. This decision support requires knowledge so that automatic decisions about diagnosis can be made. It might be acquired through the process of knowledge discovery in databases where patient records are utilized. It is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [6]. Its computational stage is called data mining. There are several recent papers which use data mining techniques in relation to heart disease such as decision tree, Naive Bayes, nearest neighbor and neural network [1], [4], [13], [16], [24]. In [1], [4], [16] and [24], various existing data mining techniques are surveyed and compared experimentally for the purpose of heart disease diagnosis. Measures sensitivity, specificity and ROC graph are employed for comparisons in publications [16] and [24]. Paper [13] is specialized on a multilayer neural network which is trained with backpropagation and simulated on feedforward network. A probability approach with Naive Bayes and its improvement with a presented supervised discretization of numerical data are analyzed in this paper. This approach takes the specifics of given heart disease patients into consideration with the goal of increasing the accuracy of detecting the occurrence of heart disease.

The paper is organized as follows. In Section 2, the used heart data and its attributes are described and analyzed in detail. The employed discretization of the data is explained in Section 3. Validation methods and achieved experimental results are discussed in Section 4. Our conclusions are presented in Section 5.

Attribute	Data Type	Values	Units
Age (B_1)	Numerical	29 - 77	years
Sex (B_2)	Categorical	*female* $(b_{2,1})$ *male* $(b_{2,2})$	N/A
Chest Pain Type (B_3)	Categorical	*typical angina* $(b_{3,1})$ *atypical angina* $(b_{3,2})$ *nonanginal* $(b_{3,3})$ *asymptomatic* $(b_{3,4})$	N/A
Resting Blood Pressure (B_4)	Numerical	94 - 200	mmHg
Serum Cholesterol Level (B_5)	Numerical	126 - 564	mg/dL
Fasting Blood Sugar Over 120 (B_6)	Categorical	no $(b_{6,1})$ yes $(b_{6,2})$	N/A
Resting ECG Results (B_7)	Categorical	*normal* $(b_{7,1})$ *ST-T wave abnormality* $(b_{7,2})$ *left ventricular hypertrophy* $(b_{7,3})$	N/A
Maximal Achieved Heart Rate (B_8)	Numerical	71 – 202	bpm
Exercise Induced Angina (B_9)	Categorical	absent $(b_{9,1})$ present $(b_{9,2})$	N/A
ST Segment Depression (B_{10})	Numerical	0 – 6.2	mm
ST Segment Slope (B_{11})	Categorical	*upsloping* $(b_{11,1})$ *flat* $(b_{11,2})$ *downsloping* $(b_{11,3})$	N/A
Colored Vessels (B_{12})	Numerical	0, 1, 2, 3	count
Thallium Heart Scan (B_{13})	Categorical	*normal* $(b_{13,1})$ *fixed defect* $(b_{13,2})$ *reversible defect* $(b_{13,3})$	N/A
Heart Disease (D)	Categorical	*absent* (d_1) *present* (d_2)	N/A

Table 1. Attributes.

2. Statlog Heart Data

The data contains records about patients who have been subject to diagnosis of heart disease and includes their final diagnosis and physical and biochemical information about them. There are 270 patients (instances) in set $\boldsymbol{V}$ who are described by thirteen describing attributes in set $\boldsymbol{B}$ and classified into one class attribute D [8]. Attributes in the database are presented in Table 1. Describing attributes $\boldsymbol{B}$ are defined as $\boldsymbol{B}$ = $\{B_1$; …; B_k; …; $B_{13}\}$ = {*Age*; *Sex*; *Chest Pain Type*; *Resting Blood Pressure*; *Serum Cholesterol Level*; *Fasting Blood Sugar Over 120*; *Resting ECG Results*; *Maximal Achieved Heart Rate*; *Exercise Induced Angina*; *ST Segment Depression*; *ST Segment Slope*; *Colored Vessels*; *Thallium Heart Scan*}. Numerical attributes B_k can have a numerical value v for a patient $\boldsymbol{p} \in \boldsymbol{V}$ and it is marked as $B_k(\boldsymbol{p}) = v$. If set $\boldsymbol{P}$ contains possible numerical values for attribute B_k, it is denoted by $B_k = \boldsymbol{P}$. Categorical attributes B_k can have a categorical value $b_{k,l}$ for a patient $\boldsymbol{p} \in \boldsymbol{V}$ and it is denoted by $B_k(\boldsymbol{p}) = b_{k,l}$. If possible categorical values are $b_{k,1}$, …, $b_{k,l}$, …, b_{k,l_k} for attribute B_k, it is denoted by $B_k = \{b_{k,1}$; …; $b_{k,l}$; …; $b_{k,l_k}\}$. Class attribute *Heart Disease* (D) classifies the patient into someone that is either with heart disease or without it. The attribute can have two possible values (absent/present), i.e. D = $\{d_1$; $d_2\}$ = {*absent*; *present*}. The value of attribute D for patient $\boldsymbol{p} \in \boldsymbol{V}$ is marked as $D(\boldsymbol{p})$.

Attribute	Value	Frequency	Median	Mode
B_1	N/A	N/A	55	54
B_2	$b_{2,1}$	87	N/A	*male*
	$b_{2,2}$	183		
B_3	$b_{3,1}$	20	N/A	*asymptomatic*
B_4	N/A	N/A	130	120
B_5	N/A	N/A	245	234
B_6	$b_{6,1}$	230	N/A	*no*
	$b_{6,2}$	40		
B_7	$b_{7,1}$	131	N/A	*left ventricular hypertrophy*
	$b_{7,2}$	2		
	$b_{7,3}$	137		
B_8	N/A	N/A	153.5	162
B_9	$b_{9,1}$	181	N/A	*absent*
	$b_{9,2}$	89		
B_{10}	N/A	N/A	0.8	0
B_{11}	$b_{11,1}$	130	N/A	*upsloping*
	$b_{11,2}$	122		
	$b_{11,3}$	18		
B_{12}	N/A	N/A	0	0
B_{13}	$b_{13,1}$	152	N/A	*normal*
	$b_{13,2}$	14		
	$b_{13,3}$	104		
D	d_1	150	N/A	*absent*
	d_2	120		

Table 2. Descriptive statistics.

Descriptive statistics of the Statlog heart data is presented in Table 2. It has the frequencies of all categorical values and the medians and modes for all particular attributes. The number of patients with absent/present heart disease is 150/120. Categorical value *male* ($b_{2,2}$) for attribute *Sex*, *asymptomatic* ($b_{3,4}$) for attribute *Chest Pain Type*, *no* ($b_{6,1}$) for *Fasting Blood Sugar Over 120* and *absent* ($b_{9,1}$) for *Exercise Induced Angina* are a lot frequenter than others. There are no missing values.

3. Employed Algorithm

3.1. Algorithm of Naive Bayes

It is supposed for the algorithm that previously known heart disease patients $\boldsymbol{p} \in \boldsymbol{V}$ are described by attributes $\boldsymbol{B} = \{B_1; \ldots; B_k; \ldots; B_{13}\}$ with known values and diagnosed to class attribute $D = \{d_1; d_2\}$ as they are defined in Section 2. The algorithm is based on Naive Bayes which is a probabilistic approach that gives the probabilities that a patient $\boldsymbol{p}$ should be diagnosed as d_1 and d_2 as its output, i.e. it diagnosis a given patient automatically. The probabilities are computed according to the expression in (1):

$$p(D|\boldsymbol{B}) = \frac{p(D)\prod_{B_k\in\boldsymbol{B}} p(B_k|D)}{p(\boldsymbol{B})} \tag{1}$$

where probability $p(\boldsymbol{B})$ is computed as follows:

$$p(\boldsymbol{B}) = \sum_{d_j\in D} p(d_j) \prod_{B_k\in\boldsymbol{B}} p(B_k|d_j). \tag{2}$$

Because the expression in (2) is always some constant value and only relative values of probabilities are important, probabilities for particular $d_j \in D$ are calculated in the following way:

$$p(d_j|\boldsymbol{B}) = p(d_j) \prod_{B_k\in\boldsymbol{B}} p(B_k|d_j). \tag{3}$$

The class from D with the highest probability is considered to be the class for $\boldsymbol{p}$. If the probabilities are the same, d_2 is chosen automatically. Conditional probabilities are computed with counting occurrences of particular categorical values $b_{k,l} \in B_k$ for categorical attributes and numerical attributes are discretized at first. However, the occurrence of some value of an attribute might be zero in some cases and this would lead to the zero result of the whole expression in (3). This is not desirable as it would cause that the weight of all other attributes would not be taken into consideration. This is avoided with the Laplace estimator [19] which counts the occurrences of all possible values $b_{k,l} \in B_k$ belonging to one $d_j \in D$ and adds value one to each obtained value, the total number of occurrences is increased with the cardinality of B_k, and the conditional probabilities are recomputed with obtained values.

3.2. Indicators of the Naive Bayes Algorithm

The accuracy of diagnosis with the algorithm is influenced significantly by input data about known heart disease patients $\boldsymbol{p} \in \boldsymbol{V}$. Important indicators for Naive Bayes are probability distributions of attributes, conditional probabilities given particular output classes and correlations between attributes [17]. It is practical to look at the difference between the number of patients belonging to a value of a categorical attribute for various output classes. There are significant differences for output classes in the case of categorical attributes B_3, B_9 and B_{13} as shown in Figure 1, Figure 2 and Figure 3. Histograms of numerical attributes B_{10} and B_{12} contain noticeable extremes as demonstrated in Figure 4 and Figure 5, especially for $D = d_1$. The correlations between D and particular numerical attributes in $\boldsymbol{B}$ are in Table 3 where the strongest ones are B_8, B_{10} and B_{12}. These also have strong correlations with some other attributes and so a suitable discretization can break inter-attribute correlations.

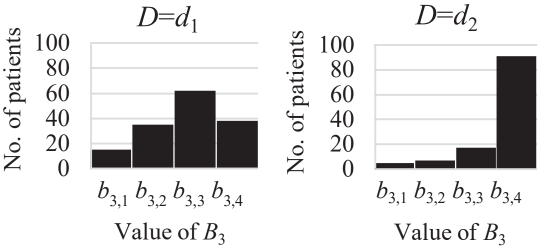

Figure 1. Bar graph of B_3 given D.

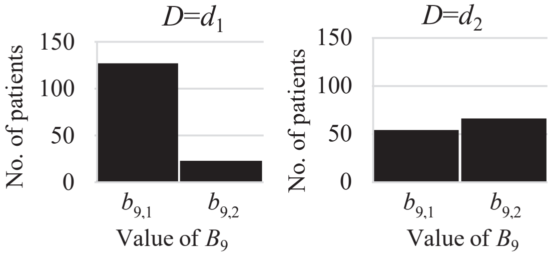

Figure 2. Bar graph of B_9 given D.

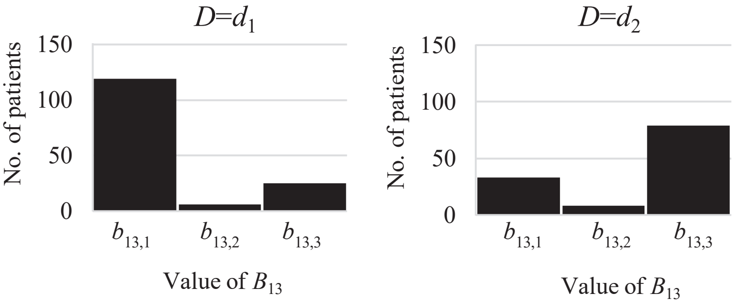

Figure 3. Bar graph of B_{13} given D.

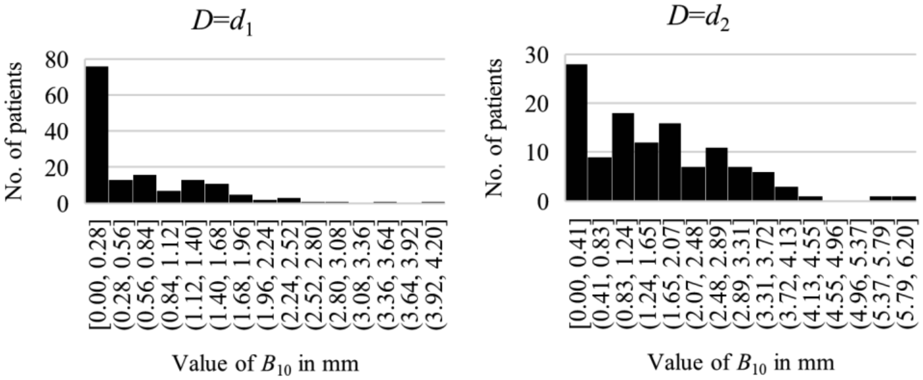

Figure 4. Histogram of B_{10} given D.

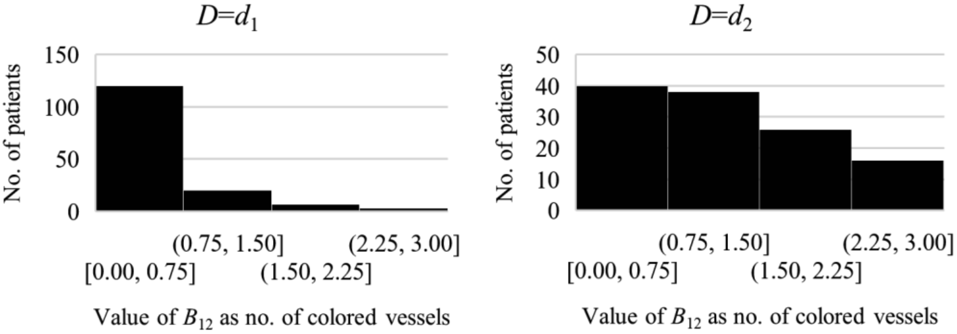

Figure 5. Histogram of B_{12} given D.

	$\boldsymbol{B_1}$	$\boldsymbol{B_4}$	$\boldsymbol{B_5}$	$\boldsymbol{B_8}$	$\boldsymbol{B_{10}}$	$\boldsymbol{B_{12}}$	$\boldsymbol{D}$
$\boldsymbol{B_1}$	1.0	0.3	0.2	-0.4	0.2	0.4	0.2
$\boldsymbol{B_4}$	0.3	1.0	0.2	-0.0	0.2	0.1	0.2
$\boldsymbol{B_5}$	0.2	0.2	1.0	-0.0	0.0	0.1	0.1
$\boldsymbol{B_8}$	-0.4	-0.0	-0.0	1.0	-0.3	-0.3	-0.4
$\boldsymbol{B_{10}}$	0.2	0.2	0.0	-0.3	1.0	0.3	0.4
$\boldsymbol{B_{12}}$	0.4	0.1	0.1	-0.3	0.3	1.0	0.5
$\boldsymbol{D}$	0.2	0.2	0.1	-0.4	0.4	0.5	1.0

Table 3. Correlations between attributes.

3.3. Discretization of Numerical Attributes

Numerical attributes are discretized with a supervised discretization which is presented here. It makes use of entropy [7] and Equal Frequency Discretization [19] with appropriate modifications and parametrizations so that the specifics of heart disease patients and the knowledge discovered during data analysis are taken into consideration. It has the following steps for an attribute $B_k \in \boldsymbol{B}$ and D:

1. Create some initial division of the values for B_k on the basis of the heart disease patients. This should be relevant medically and the division points form some starting points of the discretization. The points are marked as $R_k = \{r_1; r_2; \ldots; r_h; \ldots; r_o\}$, i.e. there are $o - 1$ discretization subintervals.
2. Sort available data for B_k and remember particular $d_j \in D$ for each data point.
3. Compute the maximal radius for the movement of the discretization point as $r = \frac{\mathrm{argmin}_{\mathrm{all}\ h,i,h \neq i} |r_h - r_i|}{3}$ and set the smallest optimization step as $\alpha = \frac{r}{n}$ where n = 1, 2, 3, … is named softness. Each $r_h \in R_k$ can move in both directions with multiplications of n to the maximal distance of αn (side points do not move).
4. Mark the number of possible mutual positions of points in R_k as m and do the following for all of these positions. Compute the entropy for each subinterval and find the average entropy, i. e. calculate expression $E_i = \frac{\boldsymbol{1}}{\boldsymbol{o-1}} \sum_{\boldsymbol{1}}^{\boldsymbol{o-1}} - \sum_{d_j \in D} p(d_j) \log_2(p(d_j))$, i = 1, 2, 3, …, m.
5. Choose the discretization which has the smallest average entropy (if there are more discretizations with the smallest entropy, choose any randomly) and transform the numerical attribute into a categorical one with this discretization.

Numerical attributes in the considered data are *Age* (B_1), *Resting Blood Pressure* (B_4), *Serum Cholesterol Level* (B_5), *Maximal Achieved Heart Rate* (B_8), *ST Segment Depression* (B_{10}) and *Colored Vessels* (B_{12}). Useful information about these attributes can be taken from medical practice. For example, serum cholesterol level is desirable if it is less than 200 mg/dL, borderline high if it is less than 240 mg/dL and high if it is 240 mg/dL and above [12]. On the basis of the domain knowledge about heart disease, division points in step 1 could lead to the following set initial discretization intervals. Attribute B_1: [0; 50), [50; ∞), attribute B_4: [0; 120), [120; 160), [160; ∞), attribute B_5: [0; 200), [200; 240), [240; ∞), attribute B_8: [0; 100), [100; 120), [120; 140), [140; ∞), B_{10}: [0; 2.5); [2.5; ∞) and attribute B_{12}: [0; 1), [1; ∞). After the execution of the remaining steps of the above algorithm for all mentioned numerical attributes and $n =$ 2, the following final discretization intervals are determined. Attribute B_1: [0; 45), [45; ∞), attribute B_4: [0; 107), [107; 165), [165; ∞), attribute B_5: [0; 206), [206; 216), [216; 245), [245; ∞), attribute B_8: [0; 106), [106; 114), [114; 146), [146; ∞), attribute B_{10}: [0; 2.5), [2.5; ∞) and attribute B_{12}: [0; 1), [1; ∞).

4. Experimental Results

The experiments showing the performance of the algorithm presented in Section 3 are described here and it is compared with other well-known data mining techniques as well. The algorithm consisting of Naive Bayes, discretization and validation tools is implemented in the Java programming language. The other known data mining techniques are implemented in Waikato Environment for Knowledge Analysis [19]. The basis of the validation is *K*-fold cross-validation where *K*=10 and measures such as sensitivity, specificity and their sum are computed. *K*-fold cross-validation partitions the heart disease data into *K* subgroups with equal numbers of cases with present and absent heart disease [9]. One subgroup is used for validation while the others are used for training and this is repeated *K* times with each of the subgroups used for validation exactly once. Sensitivity represents the ratio of patients with present heart disease who are accurately considered as the ones with heart disease. Sensitivity is computed with expression $\frac{\mathrm{tp}}{\mathrm{tp}+\mathrm{fn}}$ where tp is the number of patients who have heart disease and who are diagnosed with heart disease and fn is the number of patients who have heart disease and who are diagnosed with no heart disease. Specificity represents the amount of patients with no heart disease who are accurately considered as patients without heart disease. Low sensitivity is associated with many heart disease patients without treatment (i.e. with life-threatening states) and low specificity is associated with useless treatment of people without heart disease (i.e. with overpriced states). The sum of sensitivity and specificity takes both life-threatening and overpriced states into consideration.

	Sensitivity	Specificity	Sum
NB-Mod	**0.900**	**0.842**	**1.742**
NB	0.840	0.817	1.657
MLP	0.880	0.800	1.680
DT	0.840	0.692	1.532
NN	0.773	0.717	1.490

Table 4. Experimental results of particular techniques.

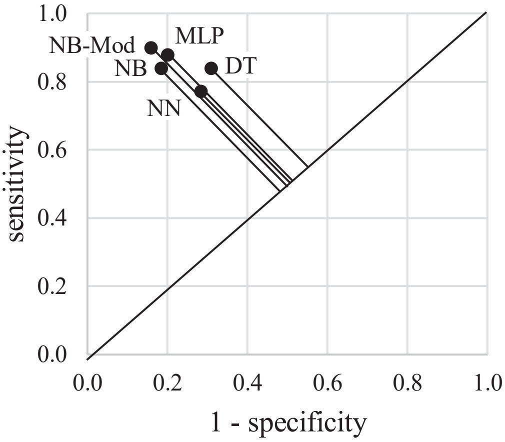

Figure 6. ROC graph.

The results of experiments are shown in Table 4 where the rows correspond to various algorithms and the columns are associated with computed measures. NB-Mod represents the method which is described in Section 3 and uses Naive Bayes with given supervised discretization which takes expert domain knowledge about heart disease patients into consideration. NB is Naive Bayes which is implemented in Waikato Environment for Knowledge Analysis as class NaiveBayes and employs well-known Fayyad-Irani's discretization [5]. MLP, DT and NN are a neural network using multilayer perception, decision tree C4.5 and a nearest neighbor classifier implemented in Waikato Environment for Knowledge Analysis as classes MultilayerPerceptron, J48 and IBk, respectively. The best results are shown in bold and they are achieved by presented NB-Mod whose sum of sensitivity and specificity equals 1.742 since a higher value of the sum is considered to be better. The ROC curve in Figure 6 shows the results in a way which is typical for medical experts.

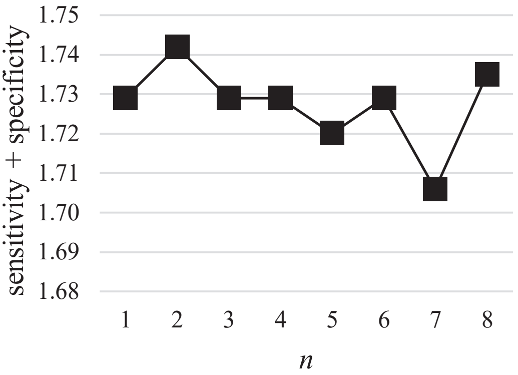

Figure 7. Accuracy results for various values of softness *n*.

Attribute	Sum for the Attribute Removal	Sum for One Attribute
B_1	1.726	1.066
B_2	1.686	1.280
B_3	1.632	1.506
B_4	1.712	0.984
B_5	1.712	1.076
B_6	1.736	0.882
B_7	1.720	1.176
B_8	1.686	1.300
B_9	1.690	1.396
B_{10}	1.668	1.160
B_{11}	1.688	1.386
B_{12}	1.582	1.466
B_{13}	1.598	1.518

Table 8. The impact of individual attributes for two boundary selections of attributes.

The sum of sensitivity and specificity for several n representing the softness as described in Section 3.3 is depicted in Figure 7. It is clear that the algorithm gives stable improvements for various initial settings. In Table 8, the sum of sensitivity and specificity are computed for two boundary selections of attributes which help analyze the results with the employed discretization. The second column shows the value of the sum when the attribute in the same row is not used and all the others are. It seems that the removal of attribute *Colored Vessels* (B_{12}) with 1.582 or *Thallium Heart Scan* (B_{13}) with 1.598 would lead to significant reduction of the sum. On the other hand, it would not have much impact if *Fasting Blood Sugar Over 120* (B_6) with 1.736 was not used at all. The third column shows the sum when only the attribute in the same row is used for training. The strongest attributes for the sum with the employed algorithm are *Chest Pain Type* (B_3) with 1.506, *Exercise Induced Angina* (B_9) with 1.396, *Colored Vessels* (B_{12}) with 1.466 and *Thallium Heart Scan* (B_{13}) with 1.518. These attributes are also strong according to the indicators used in Section 3.2 and so in the data there is no substantial distortion weakening the strength of these attributes.

5. Conclusions

A decision support tool for automatic diagnosis of heart disease with a Naive Bayes classifier using a supervised discretization which takes domain expert knowledge into consideration was presented. The discretization method took initial divisions of values from experts and used them to produce the final discretization of numerical attributes. The tool is meant to support early recognition of heart disease as precisely as possible since this may be difficult due to situations with no obvious symptoms and it is important to start the treatment early without necessary costs. Statlog heart data was used for validation of the tool through 10-fold cross-validation with measures sensitivity, specificity and their sum. Sensitivity is associated with life-threatening states, specificity with costs and their sum combines both. The higher the sum is, the better the result is. With the incorporated discretization, sensitivity achieved 0.900, specificity 0.842 and their sum 1.742, which was better than well-known Naive Bayes with Fayyad-Irani's discretization, a neural network using multilayer perception, decision tree C4.5 and a nearest neighbor classifier. Overall, the results indicated the incorporated discretization was useful for more precise recognition of heart disease.

Acknowledgements

The authors would also like to thank Roman Trisliar for implementations of algorithms and experiments on the Statlog heart data as a part of his studies leading to the completion of his thesis under the supervision of Jan Bohacik.

References

[1] S. Banu, S. Swamy, “Prediction of heart disease at early stage using data mining and big data analytics: A survey,” in International Conference on

Electrical, Electronics, Communication, Computer and Optimization Techniques, 2016, pp. 256 – 261. Available: http://ieeexplore.ieee.org/abstract/document/7955226/

[2] A. D. K. Bowry, J. Lewey, S. B. Dugani, N. K. Choudhry, "The burden of cardiovascular disease in low- and middle-income countries: Epidemiology and Management," Canadian Journal of Cardiology, vol. 31, no. 9, pp. 1151-1159, 2015. Available: https://www.sciencedirect.com/science/article/pii/S0828282X15005073?via%3Dihub

[3] F. P. Cappuccio, M. A. Miller, "Cardiovascular disease and hypertension in sub-Saharan Africa: burden, risk and interventions," Internal and Emergency Medicine, vol. 11, pp. 299-305, 2016. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4820479/pdf/11739_2016_Article_1423.pdf

[4] S. Ekiz, P. Erdogmus, "Comparative study of heart disease classification," in Electric Electronics, Computer Science, Biomedical Engineerings' Meeting, 2017. Available: http://ieeexplore.ieee.org/document/7956761/

[5] U. M. Fayyad, K. B. Irani, "Multi-Interval discretization of continuous-valued attributes for classification learning," in International Joint Conference on Uncertainty in AI, 1993, pp. 1022-1027. Available: https://trs.jpl.nasa.gov/bitstream/handle/2014/35171/93-0738.pdf?sequence=1&isAllowed=y

[6] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, "From data mining to knowledge discovery in databases," AI Magazine, vol. 17, no. 3, pp. 37-54, 1996. Available: https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1230/1131

[7] R. Kohavi, J. Dougherty, M. Sahami, "Supervised and unsupervised discretization of continuous features," in International Conference on Machine Learning, 1995, pp. 194-202. Available: http://robotics.stanford.edu/users/sahami/papers-dir/disc.pdf

[8] Lichman, M. UCI Machine Learning Repository. Irvine, CA, USA: University of California, School of Information and Computer Science, 2013. Available: http://archive.ics.uci.edu/ml

[9] G. J. McLachlan, K.-A. Do, C. Ambroise, Analyzing Microarray Gene Expression Data. San Diego, USA: Willey, 2004. Available: http://onlinelibrary.wiley.com/book/10.1002/047172842X

[10] National Heart Foundation of Australia, Heart Information: Coronary Heart Disease. Australia: National Heart Foundation of Australia, 2013.

Available: https://www.heartfoundation.org.au/images/uploads/publications/CON-093-v3_Coronary_heart_disease_WEB.PDF

[11] National Center for Chronic Disease Prevention and Health Promotion, Know the Facts About Heart Disease. : National Center for Chronic Disease Prevention and Health Promotion, 2013. Available: https://www.cdc.gov/heartdisease/docs/consumered_heartdisease.pdf

[12] National Institutes of Health, “Cholesterol levels: What you need to know,” NIH MedlinePlus, vol. 7, no. 2, pp. 6-7, 2012. Available: https://medlineplus.gov/magazine/issues/summer12/articles/summer12pg6-7.html

[13] E. O. Olaniyi, O. K. Oyedotun, A. Helwan, “Neural network diagnosis of heart disease,” in International Conference on Advances in Biomedical Engineering, 2015, pp. 21-24. Available: http://ieeexplore.ieee.org/document/7323241/

[14] S. Palaniappan, R. Awang, “Intelligent heart disease prediction system using data mining techniques,” in International Conference on Computer Systems and Applications, Doha, Qatar, 2008. Available: http://ieeexplore.ieee.org/document/4493524/

[15] M. Pignone, A. Fowler-Brown, M. Pletcher, J. A. Tice, “Screening for asymptomatic coronary artery disease: A systematic review for the U.S. Preventive Services Task Force,” Systematic Evidence Review, no. 22, 2003. Available: https://www.ahrq.gov/downloads/pub/prevent/pdfser/chdser.pdf

[16] M. Sultana, A. Haider, M. S. Uddin, “Analysis of data mining techniques for heart disease prediction,” in International Conference on Electrical Engineering and Information Communication Technology, 2016. Available: http://ieeexplore.ieee.org/document/7873142/

[17] S. Taheri, M. Mammadov, A. M. Bagirov, “Improving Naive Bayes classifier using conditional probabilities”, in Australian Data Mining Conference, 2011, pp. 63-68. Available: https://dl.acm.org/citation.cfm?id=2483637

[18] Victorian Government, Fainting and Collapse. Melbourne, Australia: Victorian Government, 2011. Available: http://www.health.vic.gov.au/edfactsheets/downloads/fainting-and-collapse.pdf

[19] I. H. Witten, E. Frank, M. A. Hall, Practical Machine Learning Tools and Techniques (3rd Edition). Burlington, MA, USA: Morgan Kaufman Publishers, 2011. Available: https://www.sciencedirect.com/science/book/9780123748560

[20] World Health Organization. Global Status Report on Non-Communicable Diseases 2010. Geneva: World Health Organization, 2011. Available: http://apps.who.int/iris/bitstream/10665/44579/1/9789240686458_eng.pdf

[21] World Health Organization, World Heart Federation, World Stroke Organization. Global Atlas on Cardiovascular Disease Prevention and Control. : World Health Organization, 2011. Available: http://apps.who.int/iris/bitstream/10665/44701/1/9789241564373_eng.pdf

[22] R. Wu, W. Peters, M. W. Morgan, "The next generation clinical decision support: Linking evidence to best practice," Journal Healthcare Information Management, vol. 16, no. 4, pp. 50-55, 2002. Available: http://www.himss.org/jhim/archive/volume-16-number-4-2002

[23] Z. M. Zhang, P. M. Rautaharju, R. J. Prineas, C. J. Rodriguez, L. Loehr, W. D. Rosamond, D. Kitzman, D. Couper, E. Z. Soliman, “Race and sex differences in the incidence and prognostic significance of silent myocardial infarction in the atherosclerosis risk in communities (ARIC) study,” Circulation, vol. 133, no. 22, pp. 2141-2148, 2016. Available: http://circ.ahajournals.org/content/133/22/2141.long

[24] I. A. Zriqat, A. M. Altamimi, M. Azzeh, “A comparative study for predicting heart diseases using data mining classification methods,” International Journal of Computer Science and Information Security, vol. 14, no. 12, pp. 868-879, 2016. Available: https://arxiv.org/ftp/arxiv/papers/1704/1704.02799.pdf