Knowledge and Information Systems (2020) 62:1497–1510
https://doi.org/10.1007/s10115-019-01391-w

REGULAR PAPER

Effective construction of classifiers with the k-NN method supported by a concept ontology

Jan Bazan[1] · Stanisława Bazan-Socha[2] · Marcin Ochab[1] · Sylwia Buregwa-Czuma[1] · Tomasz Nowakowski[3] · Mirosław Woźniak[4]

Received: 10 July 2018 / Revised: 3 August 2019 / Accepted: 5 August 2019 / Published online: 21 August 2019

Abstract
In analysing sensor data, it usually proves beneficial to use domain knowledge in the classification process in order to narrow down the search space of relevant features. However, it is often not effective when decision trees or the k-NN method is used. Therefore, the authors herein propose to build an appropriate concept ontology based on expert knowledge. The use of an ontology-based metric enables mutual similarity to be determined between objects covered by respective concept ontology, taking into consideration interrelations of features at various levels of abstraction. Using a set of medical data collected with the Holter method, it is shown that predicting coronary disease with the use of the approach proposed is much more accurate than in the case of not only the k-NN method using classical metrics, but also most other known classifiers. It is also proved in this paper that the expert determination of appropriate structure of ontology is of key importance, while subsequent selection of appropriate weights can be automated.

Keywords k-nearest neighbour algorithm · Ontology similarity metrics · Holter measurement · Coronary disease

1 Introduction

The approximation of complex concepts using exclusively sensor data sets often proves difficult, owing to the intricate nature of real-world processes, presence of direct and indirect relations and interactions between objects involved in those processes. Numerous concepts

✉ Jan Bazan
bazan@ur.edu.pl

1 Interdisciplinary Centre for Computational Modelling, University of Rzeszow, Pigonia 1, 35-310 Rzeszow, Poland

2 Department of Internal Medicine, Faculty of Medicine, Jagiellonian University Medical College, Skawinska 8, 31-066 Kraków, Poland

3 Department of Angiology, Jagiellonian University Medical College, Skawinska 8, 31-066 Kraków, Poland

4 Angiomed Private Medical Centre, Skawinska 8, 31-066 Kraków, Poland

have been developed of using domain knowledge in classifier construction with a view to taking these phenomena into account. Domain knowledge is predominantly used to narrow down the search space and facilitate interpretation of results. Such knowledge is thus used mainly in data preparation in order to eliminate irrelevant features, select the most valuable ones or develop new derivative features. The literature records a material favourable effect of such use of domain knowledge on the performance of certain data exploration methods. For instance, in Sinha and Zhao [29] and Zhao, Sinha and Ge [36] the effect is analysed of using the knowledge on the efficiency of the following classifiers: logistic regression, artificial neural networks, the k-NN (k-nearest neighbour) method, naive Bayes classifiers, decision trees and the SVM (support vector machine) method. It has been observed, though, that the use of domain knowledge for selection purposes proved least efficient in the event of decision trees and the k-NN method.

For such methods as k-NN, it is of key importance to evaluate the distance between—or in other words—similarity of two objects (e.g. patients). This requires data to be analysed at numerous levels of abstraction. However, given the semantic distance of complex concepts from sensor data, this is not feasible for classic modelling methods based on features being measured. Thus, the definition of such metrics (distance measures) remains a major challenge in data exploration. There do exist methods of defining similarity relation by way of building a metric (distance function) based on simple strategies of aggregating local similarities of the objects being compared (see Bazan [4] for more details). A chosen distance formula is there optimized by tuning local similarity features and parameters used to aggregate them. However, the main challenge in approximating the metric is the selection of such local similarities and a way of aggregation thereof, while domain knowledge shows that there usually are numerous various aspects of similarity of elements being compared. Each aspect should be examined specifically, in line with the domain knowledge. Further, also the aggregation of various aspects into a global similarity or distance should be done based on the knowledge. Therefore, the authors propose to define a semantic metric (for measuring the distance between objects) founded on a concept ontology (based on the domain knowledge) and to use it for the k-NN classifier. Ontology is understood as a finite set of concepts arranged in a hierarchy equipped with relations between concepts from different hierarchy levels.

2 Related work

For a review of existing approaches to measuring distance between concepts, the reader is referred to Pedersen et al. [25] and Taieb et al. [31]. Measures of a semantic similarity and a kinship were divided there into such types as: based on paths in a concept ontology, based on information content and context vectors. Rada et al. [27] define the notion of the semantic distance as the length of the shortest path connecting two concepts in the ontology. The longer the path, the more semantically the concepts are away. The measure of the semantic similarity between concepts based on the length and depth of the path was proposed by Wu and Palmer [35]. This approach uses the number of "is-a" edges from concepts to the nearest common LCS (lowest common subsumer) and the number of edges to the root of taxonomy. Leacock and Chodorow [21] proposed a measure of semantic similarity based on the shortest path in the lexical WordNet database [34]. The path length is scaled using the maximum taxonomy depth to a value between 0 and 1, and the similarity is calculated as the negative logarithm of this value. A measure of similarity based on the concept of information content (IC) was presented by Resnik [28]. IC, which is a measure of the specificity of a concept,

is calculated for each concept in the hierarchy based on the frequency of occurrence of this concept in a broader context. Using the concept of IC, Resnik proposes a measure in which the semantic similarity of two concepts is proportional to the amount of information they share. Lin et al. [22] proposed extending Resnik's work by scaling the information content of the superior concept of LCS by the information content of individual concepts. Hsu et al. [20] even provided a representation of such distance in the form of distance hierarchy enhancing concept classification by assigning weights to inter-concept links (edges). The distance between two values of a (categorical or numeric) feature is there measured as the total weight of edges along the path connecting two nodes (concepts), with the weights defined by an expert based on the domain knowledge. All methods of measuring semantic distance with the use of domain knowledge, described in the above-mentioned articles, relate to the comparison of concepts or values of features, which makes them useful in, for instance, discretization of features. The author-proposed method of constructing an ontology-based metric uses another approach. It is designed to determine similarity of objects covered by respective denotations of concepts, and not of concepts themselves or their features.

3 Construction of a classifier

The similarity function proposed for the purposes of exploration of a set of actual medical data enables patients to be compared in terms of the acuteness of coronary disease and thus to be evaluated for the risk of health- and life-threatening consequences. The more acute the disease, the greater the risk of heart incidents [dangerous rhythm disturbances, acute myocardial ischaemia or sudden cardiac death (SCD)]. Experimental data were provided by the Second Department of Internal Medicine of the Jagiellonian University Medical College. Two data sets were collected containing ECGs recorded with the Holter method and supplemented with clinical data of patients suffering from stable myocardial ischaemia (with sinus rhythm). From the first set (HOLTER_I), 19 features of 70 patients tested in 2006–2009 with the use of Aspel's three-channel HolCARD 24W system were used. From the second set (HOLTER_II), 20 features of 200 patients tested in 2015–2016 with the use of 12-channel R12 monitor of the BTL CardioPoint-Holter H600 v2-23 system were used. Table 1 presents the key profile and angiographic data of both sets. Our research was designed to develop an efficient k-NN classifier with the use of the proposed similarity measure as the metric. The occurrence and non-occurrence of stable coronary disease (binary decision) were chosen as decision classes.

In the first stage of similarity function construction, hierarchical ontology was defined containing concepts referring to stable myocardial ischaemia. At the bottom level, sensor features (sourced directly from data set) were placed. They were selected from the entire data set so as to correspond to the recognized SCD prognostic factors [10]. Then at each level of ontology, by assigning an appropriate weight, the materiality of a given concept with respect to the higher-level concept was defined. A domain expert chose all the weights arbitrarily as a number from the (0,1) interval. The thus developed ontology is presented in Fig. 1. To benchmark prognosis efficiency, an ontology of the same structure, but with weights determined with a Monte Carlo method, was also used at the experimental stage.

The next step consisted in defining an algorithm for computing values of the function measuring similarity of objects with the use of the defined ontology with weights assigned.

The ultimate stage was the construction of a k-NN classifier using the developed metric of semantic similarity of patients.

Table 1 Clinical profile of tested populations (the HOLTER_I and HOLTER_II sets)

Feature	HOLTER_I Value	HOLTER_II Value
Number (N) of people	70	200
Age	60.6 (38–75)	68.3 (40–89)
Gender (men/women)	42 (60%)/28 (40%)	121 (60%)/79 (40%)
Diagnosed with hypertension	65 (92.9%)	168 (84%)
Myocardial infarction in medical history	14 (20%)	79 (39.5%)
Stroke in medical history	2 (2.9%)	No data
Diagnosed with lower limb atherosclerosis	7(10%)	52 (26%)
Diagnosed with diabetes	16 (22.9%)	63(31.5%)
Tobacco smokers	44 (62.9%)	57 (28.5%)
No narrowing of coronary arteries	34 (49%)	118(59%)
Diagnosed with vascular disease	36 (51%)	82 (41%)

Data given as an absolute number (and in parentheses, as per cent of global population), with the exception of age, where the average and (in parentheses) the range are given

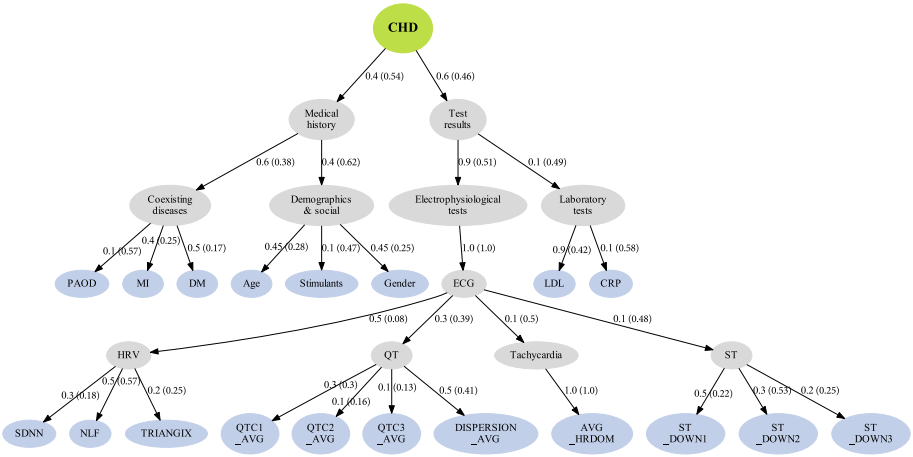

Fig. 1 CHD ontology with expert-assigned weights for the HOLTER_I set (as a comparison values in parentheses present weights determined with Monte Carlo method)

3.1 Construction of ontology

Determination of ontology-based distance requires predefining a concept ontology covered by the term which defines the decision problem. In line with the construction plan for such ontology, proposed by Noy and McGuinness [24], medical sciences were chosen as the domain and cardiology as the field. Then concepts were identified indicating the advancement of myocardial ischaemia, such as: alterations in the anamnesis, alterations in supplementary tests, epidemiological risks, coexisting diseases, alterations in electrophysiological tests or deviations in laboratory tests. These notions served a basis for defining the following ontology concepts: CHD (coronary heart disease), anamnesis, supplementary tests, epidemiology, coexisting diseases, electrophysiological tests, laboratory tests, ECG, HRV, QT, tachycardia

Table 2 Prognostic SCD factors in anamnesis

Examination subject	Risk	Description
Epidemiology	Age	Over 65 years of age
	Gender	Men:women = 4:1
	Substances	Tobacco (a risk)
Coexisting diseases	Diabetes	Prognostically adverse
	Hypertension	Prognostically adverse
	Prognostically adverse	Prognostically adverse (number of occurrences)
	Lower limb atherosclerosis	Prognostically adverse
	Stroke	Prognostically adverse

Table 3 Prognostic SCD factors in supplementary tests

Examination subject	Risk	Description
Electrophysiological tests (ECG)	ST alterations	Depression $>$ 0.5 mm; elevation $>$ 1 mm
	QTc alterations	Reference $<$ 440 ms
	Tachycardia	Resting HR $>$ 70 bpm number of occurrences
	Arrhythmia	0–5 (Lown scale), where 0 means no PVC
Laboratory tests	Troponin I	Reference $<$ 0.01 μg/l
	LDL	Reference $<$ 3.5 mmol/l
	CRP	Reference $<$ 10 mg/l

and ST. Then, using top-down approach, the concepts were arranged hierarchically into a tree-like structure, starting from the most general concept (at the top), down to the most special ones (at the bottom). Each concept was assigned a property in the form of a weight, a number in the (0,1) interval, reflecting the concept materiality with respect to the concept preceding on the tree (one level up), with the proviso that the sum of weights assigned to all successors (one level down along the tree paths) of a given concept is 1. The last stage of the construction consisted in defining instances of individual concept in the form of recognized SCD prognostic factors [1,12,14,26] corresponding to appropriate data set features. Selected risks are presented in Tables 2 and 3. In the CHD ontology thus developed (see Fig. 1), 19 risks were used, to which experts assigned weights in proportion to their relative importance in the denotation of the respective concept. At the bottom level, there are concept instances directly from the data set. The ontology proposed includes only selected concepts, present in the data sets. It may though be easily extended to include further elements. We should mention here that the literature does not specify the required number of risks: the larger the number of risk factors, the greater the risk of heart incidents [dangerous rhythm disturbances, acute myocardial ischaemia or sudden cardiac death (SCD)]. The OWL technology [3] was used to record and store the ontology developed.

It should be noted that the thresholds values which can be seen in the column called "*Description*" of mentioned tables represent the current medical knowledge, but were not used for determining values of any symbolic attributes used in the constructed ontology. The only parameters that take symbolic values are: HA, MI, DM, $stimulants$ and $gender$. They represent one of the following facts: existence of some disease, usage of stimulant or usage of gender. Such facts are organic and any thresholding of that dichotomic values is not

required. All other parameters are numeric and while determining similarity between patients were compared using formula, so it was also not required to use thresholding mechanism which could involve usage of some arbitrarily chosen threshold values.

3.2 Determination of the ontology-based distance value

Based on the ontology thus built, we can measure the distance between objects within the denotation of concepts described by the ontology. Each concept of the ontology describes differentiation among the objects considered, patients in the case in question. Metric (distance) proposed hereinafter has been designed to help answer the question "how similar (or dissimilar) are two patients diagnosed with myocardial ischaemia?".

Standard metric-based techniques use such metrics as the Euclidean distance or, more generally, the Minkowski distance (p-norm) defined by formula (1) and (2), respectively.

$$d_{\text{Euclides}}(x, y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}, \tag{1}$$

$$d_{\text{Minkowski}}(x, y) = \sqrt[p]{\sum_{i=1}^{m}|x_i - y_i|^p}, \tag{2}$$

where m is the number of conditional features in the decision table, while $x = [x_1, x_2, \ldots, x_m]$ and $y = [y_1, y_2, \ldots, y_m]$ are the values of those m features for two objects. The parameter p is a positive integer.

However, these metrics take into consideration exclusively data collected by sensors, with no regard to interrelations among concepts at the higher level of abstractness. Moreover, they can process numeric features only. On the other hand, the ontology-based metric proposed herein can handle hierarchies and meanings of the concepts described and is free from such limitations. Its computation is a multi-stage process. In the first stage, distances are computed between feature values from sensor readings, that is, at the bottom level of the ontology. In subsequent stages, at a given ontology level, the distance between two objects in the denotation of a given concept is defined, using the distances between the respective objects in the denotations of concepts subordinate to a given concept (one level down) and their respective weights measuring their impact on the higher-level concept.

The similarity function measuring distance between two objects u_i and u_j with respect to a numeric sensor-monitored feature a at the bottom ontology level is defined by formula (3) [33]:

$$d_{\text{num}}(u_i, u_j, a) = \frac{|a(u_i) - a(u_j)|}{R_a} \quad \text{for } i, j \in \{1, \ldots, n\}, \tag{3}$$

where n stands for the number of objects and R_a is the range of the feature values. The range may be defined as the difference between the greatest and the least values of the feature in a given data set or it may be known from the domain knowledge. Given a lack of accurately determined extreme values for certain SCD risks, the former approach is used herein. The similarity function measuring distance with respect to a symbolic (non-numeric) sensor-monitored feature (attribute) a is defined with the use of the value difference metric (VDM) method [30], in accordance with formula (4):

$$d_{\text{symb}}(u_i, u_j, a) = \sum_{d_c \in D} |P(dec = d_c | a(u_i) = v) - P(dec = d_c | a(u_j) = w)|, \tag{4}$$

where D stands for the set of decision classes, P is a probability distribution on the set of decision values (see formula 5), $v, w \in V_a$, which is the domain of the feature a.

$$P(dec = d_c | a(u) = v) = \frac{|\{u \in U : dec(u) = d_c \wedge a(u) = v\}|}{|\{u \in U : a(u) = v\}|}, \tag{5}$$

where U is a non-empty finite set (the "universe"), whose elements are called objects: $U = u_1, u_2, \ldots, u_n$, and $dec(u)$ is the value of the decision feature for an object u.

Finally, the similarity function defining the distance between two objects u_i and u_j with respect to a concept C arranged at a higher ontology level is defined in accordance with formula (6):

$$d_{\text{onto}}(u_i, u_j, C) = \begin{cases} \sum_{s \in S} w_s \cdot d_{\text{onto}}(u_i, u_j, C_s) & \\ \sum_{a \in A} w_s \cdot d_{\text{num}}(u_i, u_j, a) & \text{for numeric attribute a} \\ \sum_{a \in A} w_s \cdot d_{\text{symb}}(u_i, u_j, a) & \text{for symbolic attribute a} \end{cases}, \tag{6}$$

where S stands for the set of subordinate concepts lying in the denotation of the concept C (unless sensor-monitored features lie one level down), w_s stands for the weight of a given subordinate concept s or a feature (at the bottom level) and C_s represents a concept subordinate to C, one level down (at the bottom level, it is a feature a).

It is easy to prove that the proposed similarity function meets the classic three conditions known as metric axioms (the identity and symmetry axioms follow directly from the properties of the absolute value; the triangle inequality may be proved by induction) [10]. Thus, one can try to use this function in the k-NN method as a distance measure.

4 Experiments

Using the proposed ontology-based metric, experiments were performed with k-NN classifiers. For the HOLTER_I data set, the myocardial ischaemia ontology presented in Fig. 1 was used. For the purposes of testing HOLTER_II data set, the ontology was slightly modified to adapt the concepts to features available in the set (see Fig. 3). The modification was necessary, because two different ECG monitors were used to collect data for the data sets, generating slightly different parameters. The SOFA (Simple Ontology Framework API) Java library [2] was used to represent ontology models.

To compare the efficiency of the classic k-NN classifier with the one using the proposed similarity metric, four types of tests were run: E1, E2, E3 and E4, described in Tables 4 and 5 for the data sets HOLTER_I and HOLTER_II, respectively. In the experiments, the implementation of k-NN was supported by the WEKA system [15] with the authors' adaptation to the ontology-based metric. The parameter k (the number of neighbours taken into consideration) was set at 3 for the HOLTER_I set and 5 for the larger HOLTER_II set. The above values were chosen experimentally to give the best results. However, taking into account that k should be an odd number and a widely known rule of thumb [17] says that reasonable value is $k = \sqrt{n}$ (where n is a number of samples), the process of searching the optimal value was started from a $k = 7$ for the HOLTER_I set and $k = 13$ for the HOLTER_II.

The individual tests were differentiated in terms of the metric used (the Euclidean distance or the metric based on the ontology from Figs. 1 and 3, comprising 31 concepts) and of the method of determining ontology concept weights (defined by an expert or randomly

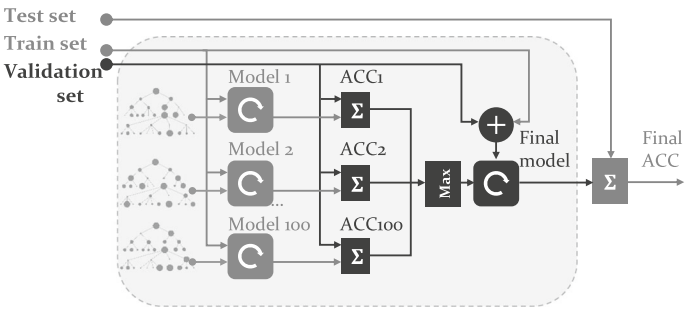

Fig. 2 Diagram of the nested cross-validation

Table 4 Results of experiments run with the use of the proposed ontology-based similarity metric for the prediction of coronary stenosis in CHD—the HOLTER_I data set

Item	E1	E2	E3	E4
Metric	Euclidean	Ontology-based	Ontology-based	Ontology-based
Notion weights	–	Expert's	Monte Carlo	Monte Carlo
Quality assessment	LOO	LOO	LOO	Nested CV
Accuracy	82.35%	94.12%	93.52% (average of 12 iterations, SD = 0.05)	98.53%

Table 5 Results of experiments run with the use of the proposed ontology-based similarity metric for the prediction of coronary stenosis in CHD—the HOLTER_II data set

Item	E1	E2	E3	E4
Metric	Euclidean	Ontology-based	Ontology-based	Ontology-based
Notion weights	–	Expert's	Monte Carlo	Monte Carlo
Quality assessment	10-CV	10-CV	10-CV	Nested CV
Accuracy	68.50%	92.50%	84.5% (average of 12 iterations, SD = 0.08)	93%

generated with a Monte Carlo method). Given the low number of items in the HOLTER_I set, the classification quality was evaluated with the n-fold cross-validation known as LOO (leaving-one-out), where the number of iterations equals the total number of objects [16,18]. For the larger HOLTER_II set, the standard tenfold cross-validation (10-CV) [9] was used, but not in the last experiment E4, where the nested cross-validation (nested CV) [32] was used. With the nested technique, external validation was performed with the LOO method (for HOLTER_I) and 10-CV (for HOLTER_II). In each train set, 100 ontology models with randomly defined weights were generated; subsequently, the highest accuracy model (ACC) was selected with the 10-CV technique for external testing. The final result is the average of all tests. Figure 2 presents the diagram of the nested cross-validation performed. The structure and results of experiments are set forth in Tables 4 and 5 for the data sets HOLTER_I and HOLTER_II, respectively.

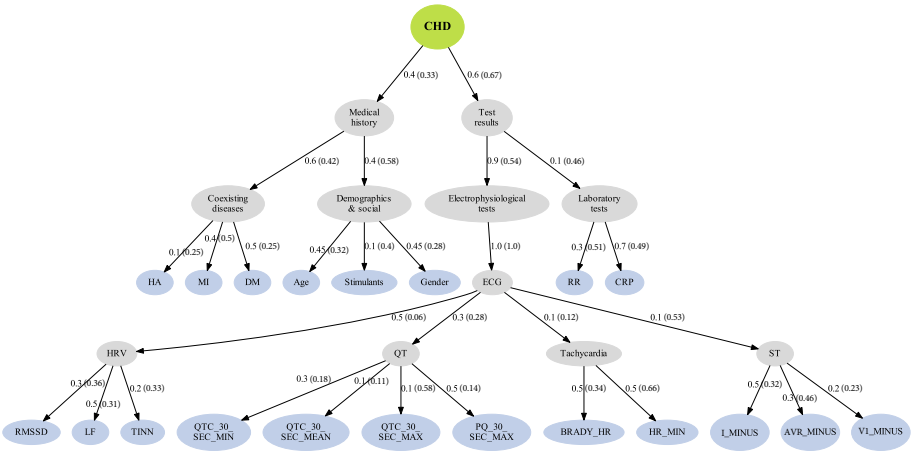

Fig. 3 CHD ontology with expert-assigned weights for the HOLTER_II set (as a comparison values in parentheses present weights determined with Monte Carlo method)

5 Conclusions

For the both data sets, the k-NN method supported by the ontology-based metric gives accuracy significantly higher than the same method supported by the Euclidean metric. For the HOLTER_I set and the ontology-based metric, the interesting thing is a minor difference in accuracy between procedure with expert-defined weights and that with randomly generated ones, which suggests that the domain knowledge-based selection of concept ontology is much more important than the selection of weights assigned to the concepts. On the other hand, the results of experiment E4 (where weights are repeatedly selected randomly and only the best ones are used in the model) indicate that, apart from the appropriate selection of concepts, appropriate weight allocation may additionally improve the accuracy of classification. The superiority of the automated (random) weight allocation is most probably attributable to difficulties faced by a human being trying to simultaneously numerically evaluate the weights with high accuracy for such a large number of ontology concepts (here 31). The exact values and differences between the weights determined by the expert and those automatically generated are shown in Figs. 1 and 3. Moreover, arbitrarily assuming 0.3 as a threshold value of the difference in the values of the weights it can be observed that for both sets, the "*electrophysiological tests*" and HRV are overestimated and the parameters "*laboratory tests*", ST and "*stimulants*" were underestimated by the expert. Therefore, the logical conclusion is a reduction in the value of "*electrophysiological tests*" at the cost of "*laboratory tests*" and the weights of HRV at the cost of ST. As seen in Tables 4 and 5, for the larger data set, the accuracy of the proposed classifier is by 24% higher than the accuracy of the classic k-NN classifier. Moreover, when compared with other classification methods examined by the authors [10], the method proposed herein is most effective in prognosis of the occurrence of material coronary stenosis in the myocardial ischaemia (see Table 6).

Given the "k-nearest neighbour" method's relatively high computational complexity, its use supported by the ontology-based metric is only feasible if classification is based on a low number of objects. However, when compared with the classic approach (using, for instance, the Euclidean distance) it proves better owing to a significant reduction in the

Table 6 Classification accuracy comparison for selected methods and the data sets HOLTER_I and HOLTER_II [10]

Method	Accuracy	
	HOLTER_I	HOLTER_II
C4.5 (WEKA) [15]	0.545	0.875
Naive Bayes (WEKA) [15]	0.394	0.560
SVM (WEKA) [15]	0.545	0.860
k-NN (WEKA) [15]	0.667	0.665
Random forest (WEKA) [15]	0.515	0.750
Multilayer perceptron (WEKA) [15]	0.548	0.825
Global discretization + all rules (RSES) [8]	0.667	–
Local discretization + all rules (RSES) [8]	0.758	0.885
ModLEM (ROSE2) [23]	0.576	0.905
CTree-Disc [5]	0.758	0.894
CTree-DiscW [6]	0.879	0.883
VTree-Disc [7]	0.909	0.930
Euclidean k-NN (E1)	0.8235	0.6850
k-NN-Ont (E2)	0.9412	0.9250
k-NN-Ont (E3)	0.9352	0.8450
k-NN-Ont (E4)	0.9853	0.9300

number of features. Namely, the ontology developed by a domain expert enabled the number of features to be reduced from 595 available in the set to just 20, thus materially shortening the computation time. Apart from computational complexity and memory requirements, another shortcoming of the method proposed herein is, as for now, a lack of a mechanism for verifying ontology quality. An interesting direction for further research also appears to be the use of the ontology-based semantic metric proposed to solving grouping problems with such tools as the c-means or hierarchical method.

Due to the fact that machine learning and especially the latest deep learning approach lack in the desired feature of explainability [13,19], we think that the presented concept ontology could be also found useful in the process of building self-explanatory artificial neural networks supported by domain knowledge.

Another interesting idea would be to use a fuzzy ontology [11]. That way some domain knowledge which is based on the threshold values (as in Tables 2 and 3) could be safely introduced to the model. Those values could be used to divide the numerical attributes by compartments, creating this way a group of symbolic values. Without the fuzzy approach, it could lead to some classification errors on the samples with values close to the borders of the compartments. However, using the fuzzy ontology there is no such a risk. For example (see Table 2), a patient who is 64 years old would be treated by model differently to a someone who is 40, even though they would be labelled with the same symbolic value of low risk. It seems that such an approach is worth attention because it can increase the scope of the domain knowledge used and thus additionally increase the accuracy of the prediction.

Acknowledgements We thank anonymous reviewers for their very useful comments and suggestions. This work was partially supported by the Centre for Innovation and Transfer of Natural Sciences and Engineering Knowledge of University of Rzeszow, Poland.

References

1. Al-Khatib SM, Yancy CW, Solis P, Becker L, Benjamin EJ, Carrillo RG, Ezekowitz JA, Fonarow GC, Kantharia BK, Kleinman M, Nichol G, Varosy PD (2017) 2016 AHA/ACC clinical performance and quality measures for prevention of sudden cardiac death: a report of the American College of Cardiology/American Heart Association Task Force on Performance Measures. Circul Cardiovasc Qual Outcomes 10(2):e000022
2. Alishevskikh A, Subbiah G (n.d.) Sofa: simple ontology framework API. http://sofa.projects.semwebcentral.org
3. Antoniou G, Van Harmelen F (2009) Web ontology language: owl. Handbook on ontologies. Springer, Berlin, pp 91–110
4. Bazan JG (2008) Hierarchical classifiers for complex spatio-temporal concepts. In: Peters JF, Skowron A, Rybinski H (eds) Transactions on rough sets IX, vol 5390. LNCS. Springer, Berlin, pp 474–750
5. Bazan JG, Bazan-Socha S, Buregwa-Czuma S, Pardel PW, Sokolowska B (2012) Prediction of coronary arteriosclerosis in stable coronary heart disease. In: International conference on information processing and management of uncertainty in knowledge-based systems. Springer, pp 550–559
6. Bazan JG, Buregwa-Czuma S, Pardel PW, Bazan-Socha S, Sokołowska B, Dziedzina S (2015) Predicting the presence of serious coronary artery disease based on 24 hour holter ecg monitoring. In: Transactions on rough sets XIX. Springer, pp 95–113
7. Bazan JG, Bazan-Socha S, Buregwa-Czuma S, Dydo L, Rzasa W, Skowron A (2016) A classifier based on a decision tree with verifying cuts.Fundamenta Informaticae 143(1–2):1–18
8. Bazan JG, Szczuka M (2005) The rough set exploration system. In: Peters JF, Skowron A (eds) Transactions on rough sets III. Springer, Berlin, pp 37–56
9. Bishop CM, Mitchell TM (2014) Pattern recognition and machine learning. Springer, Berlin
10. Buregwa-Czuma S (2017) Methods of applying domain knowledge to improve the quality of classifiers (In Polish), PhD thesis, University of Silesia in Katowice, Faculty of Computer Science and Materials Science, Katowice, Poland
11. Calegari S, Ciucci D (2007) Fuzzy ontology, fuzzy description logics and fuzzy-owl. In: Masulli F, Mitra S, Pasi G (eds) International workshop on fuzzy logic and applications. Springer, Berlin, pp 118–126
12. Ford ES, Giles WH, Mokdad AH (2004) The distribution of 10-year risk for coronary heart disease among U.S. adults. J Am Coll Cardiol 43(10):1791–1796
13. Goebel R, Chander A, Holzinger K, Lecue F, Akata Z, Stumpf S, Kieseberg P, Holzinger A (2018) Explainable AI: the new 42?, In: International cross-domain conference for machine learning and knowledge extraction. Springer, Berlin, pp. 295–303
14. Goff DC, Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, Greenland P, Lackland DT, Levy D, O'Donnell CJ, Robinson JG, Schwartz JS, Shero ST, Smith SC, Sorlie P, Stone NJ, Wilson PWF (2014) ACC/AHA guideline on the assessment of cardiovascular risk. Circulation 129(25 Suppl 2):S49–S73
15. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. SIGKDD Explor 11(1):10–18
16. Han J, Pei J, Kamber M (2011) Data mining: concepts and techniques. Elsevier, New York
17. Hassanat AB, Abbadi MA, Altarawneh GA, Alhasanat AA (2014) Solving the problem of the k parameter in the k-NN classifier using an ensemble learning approach, arXiv preprint arXiv:1409.0919
18. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer, New York
19. Holzinger A, Kieseberg P, Weippl E, Tjoa AM (2018) Current advances, trends and challenges of machine learning and knowledge extraction: From machine learning to explainable AI. In: International cross-domain conference for machine learning and knowledge extraction. Springer, Berlin, pp 1–8
20. Hsu C-C, Chen C-L, Su Y-W (2007) Hierarchical clustering of mixed data based on distance hierarchy. Inf Sci 177(20):4474–4492
21. Leacock C, Chodorow M (1998) Combining local context and wordnet similarity for word sense identification. WordNet Electr Lex Database 49(2):265–283
22. Lin D et al (1998) An information-theoretic definition of similarity. In: ICML, vol 98, Citeseer, pp 296–304

23. Napierała K, Stefanowski J (2010) Argument based generalization of modlem rule induction algorithm. In: International conference on rough sets and current trends in computing. Springer, pp 138–147
24. Noy NF, McGuinness DL (2001) Ontology development 101: a guide to creating your first ontology, Technical report, Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880
25. Pedersen T, Pakhomov SV, Patwardhan S, Chute CG (2007) Measures of semantic similarity and relatedness in the biomedical domain. J Biomed Inform 40(3):288–299
26. Priori SG, Aliot E, Blomstrom-Lundqvist C, Bossaert L, Breithardt G, Brugada P, Camm AJ, Cappato R, Cobbe SM, Mario CD, Maron BJ, McKenna WJ, Pedersen AK, Ravens U, Schwartz PJ, Trusz-Gluza M, Vardas P, Wellens HJJ, Zipes DP (2001) Task force on sudden Cardiac death of the European Society of Cardiology, Technical report, European Heart Journal
27. Rada R, Mili H, Bicknell E, Blettner M (1989) Development and application of a metric on semantic nets. IEEE Trans Syst Man Cybern 19(1):17–30
28. Resnik P (1995) Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint arXiv:cmp-lg/9511007
29. Sinha AP, Zhao H (2008) Incorporating domain knowledge into data mining classifiers: an application in indirect lending. Decis Support Syst 46(1):287–299
30. Stanfill C, Waltz D (1986) Toward memory-based reasoning. Commun ACM 29:1213–1228
31. Taieb MAH, Aouicha MB, Hamadou AB (2014) Ontology-based approach for measuring semantic similarity. Eng Appl Artif Intell 36:238–261
32. Varma S, Simon R (2006) Bias in error estimation when using cross-validation for model selection. BMC Bioinform 7(1):91
33. Wilson D, Martinez T (1997) Improved heterogeneous distance functions. J Artif Intell Res 6(1):1–34
34. WordNet: lexical database of English (n.d.). http://wordnet.princeton.edu/
35. Wu Z, Palmer M (1994) Verbs semantics and lexical selection, In: Proceedings of the 32nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, pp 133–138
36. Zhao H, Sinha AP, Ge W (2009) Effects of feature construction on classification performance: an empirical study in bank failure prediction. Expert Syst Appl 36(2):2633–2644

Jan Bazan is currently an associate professor in the Interdisciplinary Centre for Computational Modelling, University of Rzeszow, Poland. He received a M.Sc. in mathematics from Pedagogical University of Rzeszów, Poland, in 1990, Ph.D. in computer science from the Warsaw University in 1999, and also habilitation in computer science from the Institute of Computer Science Polish Academy of Sciences in 2010. His research interests include rough set theory, granular computing, knowledge discovery, data mining, decision support systems and adaptive systems, especially decision support systems for medical diagnosis and therapy.

Stanisława Bazan-Socha M.D., Ph.D., graduated from the Jagiellonian University Medical College in Krakow, Poland, in 1996. She received her Ph.D. in medical sciences degrees and habilitation in medical sciences from the Jagiellonian University Medical College in Krakow, Poland, in 2001 and 2017, respectively. She is a specialist in internal medicine, allergology and clinical immunology, and currently employed as an associate professor in Department of Internal Medicine, Jagiellonian University Medical College in Krakow, Poland. Her research is related to the pathology of asthma and autoimmune diseases, but also to the use of artificial intelligence methods in medical data analysis.

Marcin Ochab is currently an assistant professor in the Interdisciplinary Centre for Computational Modelling, University of Rzeszow, Poland. He received his M.Sc. in automatics and robotics and Ph.D. in computer science from AGH University of Science and Technology in Krakow, Poland, in 2006 and 2017, respectively. His research interests include expert systems for medical applications as well as security of network and information systems.

Sylwia Buregwa-Czuma was an assistant in the Interdisciplinary Centre for Computational Modelling, University of Rzeszow, Poland (2008–2017). She is graduated from the Jagiellonian University Medical College in Krakow, Poland. She also received her M.Sc. in computer science from Rzeszow University of Technology, Poland, and Ph.D. in computer science from University of Silesia in Katowice, Poland, in 2017. Her research interests include bioinformatics, data mining, decision support systems and adaptive systems, especially decision support systems for medical diagnosis and therapy.

Tomasz Nowakowski M.D., Ph.D., graduated from the Jagiellonian University Medical College in Krakow, Poland, in 1994. He is a specialist in internal medicine, cardiology and angiology. He is currently employed as a Senior Assistant in the Department of Angiology, Jagiellonian University Medical College in Krakow, Poland. His research is related to the pathology of cardiovascular diseases especially coronary and peripheral artery diseases.

Mirosław Woźniak M.D., Ph.D., graduated from the Jagiellonian University Medical College in Krakow, Poland, in 1994. He is a specialist in internal medicine, cardiology, angiology currently employed in the Department of Angiology and Cardiology Jagiellonian University Medical College in Krakow, Poland.