Prediction of coronary arteriosclerosis in stable coronary heart disease

Jan G. Bazan[1,3], Stanislawa Bazan-Socha[2],
Sylwia Buregwa-Czuma[1], Przemyslaw Wiktor Pardel[1,3], and Barbara Sokolowska[2]

[1] Institute of Computer Science, University of Rzeszów
Dekerta 2 Str., 35 - 030 Rzeszów, Poland
[2] II Dept. of Internal Medicine,
Jagiellonian University Medical College
Skawinska 8 Str., 31-066 Krakow, Poland
[3] Institute of Mathematics, Warsaw University
Banacha 2 Str., 02-097 Warsaw, Poland

Abstract. The aim of the study was to assess the usefulness of classification methods in recognizing cardiovascular pathology. From the medical point of view the study involves prediction of coronary arteriosclerosis presence in patient with stable angina using clinical data and electrocardiogram (ECG) Holter monitoring records. On the grounds of these findings the need for coronary interventions is determined. An approach to solving this problem has been found in the context of rough set theory and methods. Rough set theory introduced by Zdzisław Pawlak during the early 1980s provides the foundation for the construction of classifiers. From the rough set perspective, classifiers presented in the paper are based on a decision tree calculated on the basis of the local discretization method. The paper includes results of experiments that have been performed on medical data obtained from II Department of Internal Medicine, Jagiellonian University Medical College, Krakow, Poland.
Keywords: rough sets, discretization, classifiers, stable angina pectoris, morbus ischaemicus cordis, ECG Holter.

1 Introduction

Coronary heart disease (CHD) touches people all over the world and is caused by atherosclerosis, affecting coronary arteries. One of CHD's manifestation - angina pectoris is chest pain due to ischemia of the heart muscle. Acute angina, called unstable refers to acute coronary syndrome (ACS), and when its course is chronic, it is called stable. Coronary angiography enables assessment of coronary arteries anatomy and localization of stenosis, thus permits determination of therapeutic plan and prognosis. In the case of unaltered coronary flow the pharmacological treatment is applied, otherwise there is also a need for angioplasty or surgical treatment. In some cases the catheterization cannot be carried out. These cases involve health centers with limited access to diagnostic procedures or tight budget and patients with allergy to contrast medium and other contraindications to angiography. In this situation other methods assessing coronary arteries lesions are needed.

We propose application of clinical data together with electrocardiogram (ECG) Holter recordings as prospective candidate data for coronary artery stenosis prediction. The proposed method helps to determine the management of patients with stable angina, including the need for coronary intervention, without performing invasive diagnostic procedure that angiography is. It could also work as a screening tool for all patients with CHD. The problem appears considerable because angina is one of the most often cardiovascular disease (CVD) and according to WHO (World Health Organization) CVDs are leading cause of death globally and coronary heart disease kills more than 7 million people each year (see [12]).

The presented subject is concerned with substantial computation problem. It employs classifiers building for temporal data sets, where a *classifier* is an algorithm which enables us forecasting repeatedly on the basis of accumulated knowledge in new situations (see, *e.g.*, [2] for more details). Many approaches have been proposed to construct classifiers. Among them we would like to mention classical and modern statistical techniques, neural networks, decision trees, decision rules and inductive logic programming (see, *e.g.*, [2] for more details). Classifiers were constructed also for temporal data (see, *e.g.*, [2, 7] for more details). In the paper, an approach to solving problem has been found in the context of rough set theory and methods. Rough set theory introduced by Zdzisław Pawlak during the early 1980s provides the foundation for the construction of classifiers also for temporal data sets (see [16, 4, 3, 2]).

We present a method of classifier construction that is based on features aggregating time points (see, *e.g.*, [2]). The patients are characterized by parameters (sensors), measured in time points for some period, called *a time window*. In the ECG recordings context, the exemplary parameters are number of QRS complexes, ST interval elevations and lowerings or total power of the HRV spectrum. The aggregation of time points is performed by special functions called *temporal patterns* (see, *e.g.*, [2]), that are numerical characterization of values of selected sensor from the whole time window. We assume that computing temporal pattern's value uses a formula defined by an expert. Having computed patterns, a classifier is constructed approximating a temporal concept. In studied subject, the temporal concept means the presence of coronary arteriosclerosis. The classification is performed using a decision tree that is calculated on the basis of the local discretization (see, *e.g.*, [15, 4]).

To illustrate the method and to verify the effectiveness of presented classifiers, we have performed several experiments with the data sets obtained from Second Department of Internal Medicine, Collegium Medicum, Jagiellonian University, Krakow, Poland (see Section 4).

2 Stable Coronary Heart Disease

Stable angina pectoris (chronic ischaemic heart disease) is a common and disabling clinical syndrome. In the majority of European countries, 20 000–40 000 individuals of the population per million suffer from it (see [8]). Because of population ageing and increased frequency of risk factors the incidence of angina is still increasing.

2.1 Diagnosis

Diagnosis and assessment of angina involves history, physical examination, laboratory tests and specific cardiac investigations. Non-invasive standard investigations include a resting 12-lead ECG, ECG stress testing, echocardiography, ECG Holter monitoring. Invasive techniques used in coronary anatomy assessment are: coronary arteriography and intravascular ultrasound.

Holter ECG monitoring is a continuous recording of the ECG, done over a period of 24 hours or more. Holter software carries out an integrated automatic analysis process providing information about heart beat morphology, interval measurements, heart rate variability (HRV) and rhythm overview. There are numerous systems analyzing ECG recordings. They enable processing and aggregation of data by means of existing signal analyzing methods.

Coronary arteriography is a diagnostic invasive procedure that requires the percutaneous insertion of a catheter into the vessels and heart. Injected dye (contrast medium - CM) allows to identify the presence, localization and degree of stenosis in the coronary arteries. Coronary arteriography is considered a relatively safe procedure, but in same patients complications arise. Most of them are minor, of no long-term consequence. The risk of major complications is determined to be up to 2% (see [1]). Reactions to CM are relatively common, occurring in 1 to 12% of patients (see [5]) and most of them are mild. Moderate reactions occur in 1% of people receiving CM and frequently require treatment. Severe, life-threatening reactions are reported in 0,03 to 0,16% of patients, with an expected death rate of 1 to 3 per 100 000 contrast administrations (see [6]).

There are no routine noninvasive diagnostic procedures to assess coronary flow disturbances and when there is no opportunity to perform coronary angiography, alternative solutions to the problem are needed. Application of proposed method may select potential candidates for myocardial revascularization.

2.2 Treatment

The aim of CHD treatment is to prevent myocardial infarction and death. Pharmacological treatment should reduce plaque progression, stabilize plaque by reducing inflammation and by preventing thrombosis when endothelial failure or plaque rupture occurs. There are two methods of revascularization: surgical revascularization - coronary artery bypass graft (CABG) and percutaneous coronary intervention (PCI). PCI may denote balloon catheter angioplasty with or without implantation of stents or atherectomy. With appropriate management, the symptoms usually can be controlled and the prognosis improved.

3 Automated Prediction of Coronary Atherosclerosis Presence

Forecasting coronary stenosis in patients without performing angiography requires construction of classifier, which on the basis of available knowledge assigns objects (patients) to defined decision classes. Considered decision classes are: *patients with unaltered arteries who do not need invasive treatment* (decision class: *NO*) and *patients with coronary atherosclerosis who may need angioplasty* (decision class: *YES*). Classification thus permits decision making about coronary stenosis and therapy management.

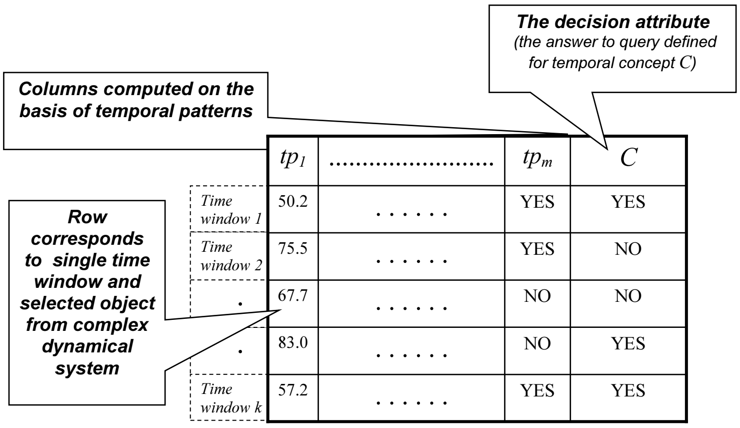
The problem of forecasting coronary atherosclerosis presence can be treated as an example of a concept approximation problem, where the term *concept* means *mental picture of a group of objects*. Such problems often can be modeled by systems of complex objects and their parts changing and interacting over time. The objects are usually linked by some dependencies, sometimes can cooperate between themselves and are able to perform flexible autonomous complex actions (operations, changes). Such systems are identified as *complex dynamical systems* or *autonomous multiagent systems* (see [2] for more details). For example, in the problem of coronary stenosis prediction, a given patient can be treated as an investigated complex dynamical system, whilst diseases of this patient are treated as complex objects changing and interacting over time.

Concepts and methods of their approximation are usually useful tools for an efficient monitoring of complex dynamic system (see [2]). Any concept can be understand as a way to represent some features of complex objects. An approximation of such concepts can be made using parameters (sensor values) registered for a given set of complex objects. However, a perception of composite features of complex objects requires observation of objects over a period called a *time window*. Such features are often represented by *temporal patterns*. In this paper, we consider temporal patterns as a numerical characterization of values of selected sensors from a time window (e.g., the minimal, maximal or mean value of a selected sensor, initial and final values of selected sensor, deviation of selected sensor values).

One can see that any temporal pattern is determined directly by values of some sensors. For example, in case of the coronary disease one can consider temporal patters such as minimal heart rate and estimated QT dispersion within a time window. We assume that any temporal pattern ought to be defined by a human expert using domain knowledge accumulated for the given complex dynamical system.

The temporal patterns can be treated as new features that can be used to approximate more complex concepts. We call them *temporal concepts*. We assume that temporal concepts are specified by a human expert. Temporal concepts are usually used in queries about the status of some objects in a particular temporal window. The approximation of temporal concepts can be defined by classifiers, which are usually constructed on the basis of decision tables. Hence, if we want to apply classifiers for approximation of temporal concepts, we have to construct a suitable decision table called a *temporal pattern table* (PT) (see Figure 1).

A temporal pattern table is constructed from a table T consisting of registered information about objects (patients) occurring in a complex dynamical system. Any row of table T represents information about parameters of a single object registered in a time window. Assume, for example, that we want to approximate a temporal concept C using table (data set) T. Initially, we construct a temporal pattern table PT as follows:


The decision attribute
(the answer to query defined for temporal concept C)
Columns computed on the basis of temporal patterns
Row corresponds to single time window and selected object from complex dynamical system

	tp_1	....	tp_m	C
Time window 1	50.2	......	YES	YES
Time window 2	75.5	......	YES	NO
.	67.7	......	NO	NO
.	83.0	......	NO	YES
Time window k	57.2	......	YES	YES

Fig. 1. The scheme of the temporal pattern table (PT)

- Construct table PT with the same objects as contained in table T.
- Any condition attribute of table PT is computed using temporal patterns defined by a human expert for the approximation of concept C,
- Values of the decision attribute (the characteristic function of concept C) are proposed by the human expert.

We assume that for any temporal pattern a formula for computing its value is given by an expert.

Next, we can construct a classifier for table PT that can approximate temporal concept C. The most popular method for classifiers construction is based on learning rules from examples (see, e.g., [16, 4, 3, 2]). Unfortunately, the decision rules constructed in this way can often be not appropriate to classify unseen cases. For instance, if we have a decision table where the number of values is high for some attributes, then there is a very low chance that a new object is recognized by rules generated directly from this table, because the attribute value vector of a new object will not match any of these rules. Therefore for decision tables with such numeric attributes some discretization strategies are built to obtain a higher quality classifiers. This problem is intensively studied and we consider discretization methods developed by Hung S. Nguyen (see [15, 4]). In this paper we use local strategy of discretization (see [4]). One of the most important notion of this strategy is the notion of *a cut*. Formally, the cut is a pair (a, c) defined for a given *decision table* $\mathbf{A} = (U, A \cup \{d\})$ in Pawlak's sense (see [16]), where $a \in A$ (A is a set of attributes or columns in the data set) and c, defines a partition of V_a into *left-hand-side* and *right-hand-side interval* (V_a is a set of values of the attribute a). In other words, any cut (a, c) is associated with a new binary attribute (feature) $f_{(a,c)} : U \to \{0, 1\}$ such that for any $u \in U$:

$$f_{(a,c)}(u) = \begin{cases} 0 & \text{if } a(u) < c \\ 1 & \text{otherwise} \end{cases} \tag{1}$$

Moreover, any cut (a, c) defines two templates, where a template we understand as a description of some set of objects. The first template defined by a cut (a, c) is a formula $T = (a(u) < c)$, while the second pattern defined by a cut (a, c) is a formula $\neg T = (a(u) \geq c)$.

In this paper, the quality of a given cut is computed as a number of objects pairs discerned by this cut and belonging to different decision classes. It is worth noticing that such quality can be computed for a given cut in $O(n)$ time, where n is the number of objects in the decision table (see, *e.g.*, [4]). The quality of cuts may be computed for any subset of a given set of objects.

In local strategy of discretization, after finding the best cut and dividing the object set into two subsets of objects (matching to both templates mentioned above for a given cut), this procedure is repeated for each object set separately until some stop condition holds. In this paper, we assume that the division stops when all objects from the current set of objects belong to the same decision class. Hence, the local strategy can be realized by using *decision tree* (see [4] and Figure 2).

The decision tree computed during local discretization can be treated as a classifier for the concept C represented by decision attribute from a given decision table $\mathbf{A}$. Let u be a new object and $\mathbf{A}(T)$ be a subtable containing all objects matching to template T defined by the cut from the current node of a given decision tree (at the beginning of algorithm work T is the template defined by the cut from the root). We classify object u starting from the root of the tree as follows:

```
Algorithm Classification by decision tree (see [4])
Step 1 If $u$ matches template $T$ found for $\mathbf{A}$
            then: go to subtree related to $\mathbf{A}(T)$
            else: go to subtree related to $\mathbf{A}(\neg T)$.
Step 2 If $u$ is at the leaf of the tree then go to 3
            else: repeat 1-2 substituting $\mathbf{A}(T)$ (or $\mathbf{A}(\neg T)$) for $\mathbf{A}$.
Step 3 Classify $u$ using decision value attached to the leaf
```

Figure 2 presents a decision tree computed for the problem of forecasting coronary atherosclerosis presence on the basis of medical data set (see Section 4).

Sample application of the tree is classification of real life objects. For example, for a patient with maximal ULF, i.e. power in ultra low frequency equal to 112 ms^2 and maximal VLF (very low frequency) equal 256, we follow from

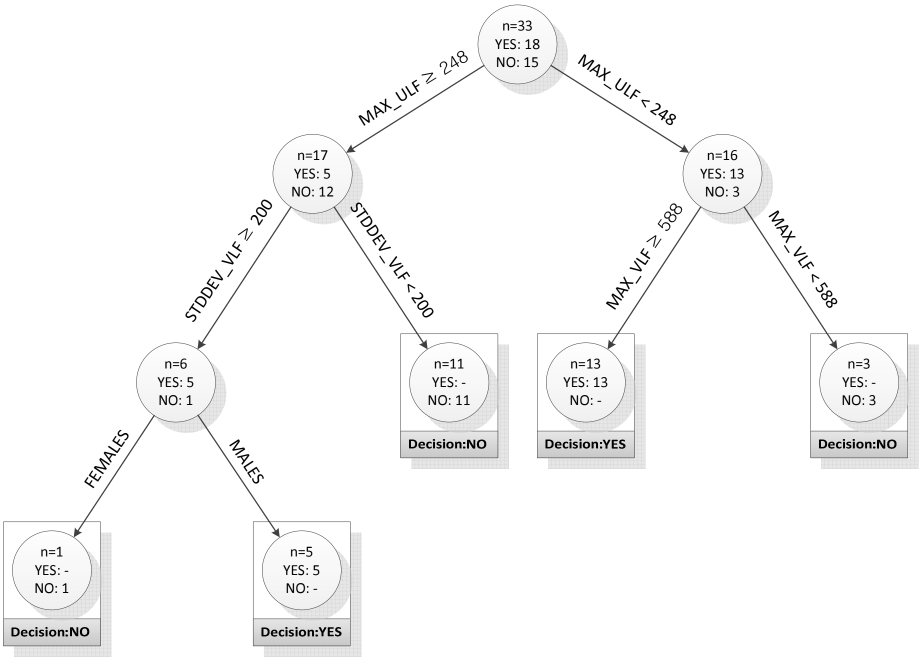


Fig. 2. The decision tree in CHD

the root of the tree, down to the right subtree, as the patient suits a template $MAX\_ULF < 248$. Then, in the next step we tread again a right tree, which consists of one node, called leaf, where we stop. The fitting path indicates that the coronary arteries of that patient are not narrowed by atherosclerosis. For a man with maximal ULF equal to 605 and standard deviation of VLF equal 509.6, we anticipate, relaying the classifier, the atherosclerotic coronary arteries stenosis presence.

4 Experimental Results

To verify the effectiveness of classifiers based on behavioral patterns, we have implemented the algorithms from the library RSH-lib, which is an extension of the RSES-lib library forming the kernel of the RSES system [3].

The experiments have been performed on the medical data set obtained from Second Department of Internal Medicine, Collegium Medicum, Jagiellonian University, Krakow, Poland. The data were collected between 2006 and 2009. Two part 48-hour Holter ECG recordings were performed using Aspel's HolCARD 24W system. There was coronary angiography after first part of Holter ECG (after first 24-hour recording). In the paper, we report results of experiments performed for the first part of Holter ECG recordings. The data set includes a detail description of clinical status (age, sex, diagnosis), coexistent diseases, pharmacological management, the laboratory tests outcomes (level of cholesterol, troponin I, LDL - low density lipoproteins) and various Holter-based indices such as: ST interval deviations, HRV, arrythmias or QT dispersion. Moreover, for Holter-based indices a data aggregation was performed resulting in points describing one hour of recording. Our group of 33 patients with normal rhythm, underwent coronary angiography and 24.2% of them required additional angioplasty, whereas 24.2% were qualified for CAGB.

All data were imported into Infobright Community Edition (ICE) environment (see [10]). ICE is an open source software solution designed to deliver a scalable data warehouse, optimized for analytic queries (data volumes up to 50TB, market-leading data compression, from 10:1 to over 40:1). After internal preprocessing in the ICE environment (e.g., a data aggregation of Holter-based indices as was mentioned above) for further processing data have been imported into Java environment.

The aim of conducted experiments was to check the effectiveness of the algorithm described in this paper in order to predict atherosclerosis in coronary arteries. Here we present the experimental results of presented method. For testing quality of classifiers we applied leave-one-out (LOO) technique, that is usually employed when the size of a given data set is small. The LOO technique involves a single object from the original data set as the validation data, and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as the validation data. As a measure of classification success (or failure) we use the following parameters well known from literature: the accuracy, the coverage, the accuracy for positive examples (Sensitivity, SN or recall), the coverage for positive examples, the precision for positive examples (Positive Predictive Value, PPV), the accuracy for negative examples (Specificity, SP), the coverage for negative examples and the precision for negative examples, also called Negative Predictive Value, NPV (see, *e.g.*, [2]).

Table 1 shows the results of applying the considered algorithm for the concept related to presence of coronary atherosclerosis in patients with stable angina.

Table 1. Results of experiments for coronary stenosis in CHD

Decision class	Accuracy	Coverage	Precision
Yes	0.778	1.0	0.778
No	0.733	1.0	0.733
All classes (Yes + No)	0.758	1.0	-

The method correctly identifies 77.8% of all patients with stenosis (SN), that's why a negative result would suggest the absence of disease. 73,3% of those who did not have stenosis (SP) were correctly identified, so a positive result means a high probability of the presence of disease. With PPV value equal

77.8%, a positive screen test is good at confirming coronary stenosis, however a negative result is also good as a screening tool at affirming that a patient does not have stenosis (NPV = 73.3%).

It is worth noticing that during LOO procedure, the most of generated trees revealed the same topology as final decision tree preserving siblings and ancestors order. The topology of the rest of trees was similar, that is, there were some differences in case of attribute values in the tree nodes and sometimes in attributes in the lower levels of generated trees. It shows that the method is quite robust to noise in data.

In Table 2 we give the results of experiments in applying other classification methods to our data. Those methods were developed in the following systems well known from literature: WEKA [17], RSES [3], and ROSE2 [14] (we used an early implementation of ModLEM algorithm [13] that is available in ROSE2). The coverage of all tested methods was equal 1.0 (every object was classified).

Table 2. Comparison results of alternative classification systems

	Accuracy			Precision	
Method	All classes	Yes	No	Yes	No
C4.5 (WEKA)	0.424	0.555	0.267	0.476	0.333
NaiveBayes (WEKA)	0.394	0.611	0.133	0.458	0.222
SVM (WEKA)	0.545	0.611	0.467	0.579	0.500
k-NN (WEKA)	0.667	0.833	0.467	0.652	0.700
RandomForest (WEKA)	0.515	0.722	0.267	0.542	0.444
Multilayer Perceptron (WEKA)	0.548	0.611	0.467	0.579	0.500
Global discretization + all rules (RSES)	0.667	0.611	0.733	0.733	0.611
Local discretization + all rules (RSES)	0.758	0.778	0.733	0.778	0.733
ModLEM (ROSE2)	?	?	?	?	?

Experimental results showed that the presented method of atherosclerosis prediction in coronary arteries gives good results and the results are comparable with results of another systems.

5 Conclusion

Presented methods were useful for development of new interesting observation and experience. We conclude that experimental outcomes showed that the proposed prediction method gives good results, also in the opinion of medical experts (compatible enough with the medical experience). But we realize that applying them in medical practice as a supporting tool for patients suffering from CHD needs clinical verification.

Acknowledgement. This work was supported by the grant N N516 077837 from the Ministry of Science and Higher Education of the Republic of Poland, the Polish National Science Centre (NCN) grant 2011/01/B/ST6/03867 and by the Polish National Centre for Research and Development (NCBiR) grant No. SP/I/1/77065/10 in frame of the the strategic scientific research and experimental development program: “Interdisciplinary System for Interactive Scientific and Scientific-Technical Information”.

References

1. ACC/AHA Guidelines for Coronary Angiography: Executive Summary and Recommendations. A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation, 1999, 99, pp. 2345–2357.
2. J. G. Bazan. Hierarchical classifiers for complex spatio-temporal concepts. In *Transactions on Rough Sets*, **IX**, LNCS 5390, 2008, pp. 474–750.
3. J. G. Bazan, M. Szczuka. The Rough Set Exploration System. In *Transactions on Rough Sets*, **III**, LNCS 3400, 2005, pp. 37–56.
4. J. G. Bazan, H. S. Nguyen, S. H. Nguyen, P. Synak, J. Wróblewski. Rough set algorithms in classification problems. In: L. Polkowski, T. Y. Lin, S. Tsumoto (Eds.), Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems, Springer-Verlag/Physica-Verlag, Heidelberg, Germany, Studies in Fuzziness and Soft Computing, vol. 56. 2000, pp. 49–88.
5. L.M. Canter. Anaphylactoid reactions to radiocontrast media. Allergy and Asthma Proceedings, 2005, 26, pp. 199–203.
6. S.T. Cochran. Anaphylactoid reactions to radiocontrast media. Current Allergy and Asthma Reports, 2005, 5, pp. 28–31.
7. A. Douzal-Chouakria, C. Amblard. Classification trees for time series. In *Pattern Recognition*, Volume 45, Issue 3, 2011, pp. 1076-1091.
8. Guidelines on the management of stable angina pectoris: executive summary. The Task Force on the Management of Stable Angina Pectoris of the European Society of Cardiology. European Heart Journal, 2006, 27, pp. 1341–1381.
9. Guidelines. Heart rate variability. Standards of measurement, physiological interpretation, and clinical use. Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology. European Heart Journal,1996, 17, pp. 354–381.
10. The Infobright Community Edition (ICE) Homepage at `http://www.infobright.org/`
11. R.A. Lange, L.D. Hillis. Diagnostic Cardiac Catheterization. Circulation, 2003, 107, pp. e111–e113.
12. J. Mackay, G.A. Mensah. The Atlas of Heart Disease and Stroke. World Health Organization, 2004.
13. K. Napierala, J. Stefanowski. Argument Based Generalization of MODLEM Rule Induction Algorithm. Proceedings of the RSCTC 2010, Springer-Verlag, Lecture Notes in Artificial Intelligence 6086, 2010, 138–147.
14. The Rough Sets Data Explorer (ROSE2) Homepage at `http://idss.cs.put.poznan.pl/site/rose.html`
15. Hung S. Nguyen. Approximate Boolean Reasoning: Foundations and Applications in Data Mining. In *Transactions on Rough Sets*, **V**, LNCS 4100, 2006, pp. 334–506.
16. Z. Pawlak, A. Skowron. Rudiments of rough sets. Information Sciences, 2007, 177, pp. 3–27.

17. The Weka 3 - Data Mining Software in Java (WEKA) Homepage at `http://www.cs.waikato.ac.nz/ml/weka/`