DOI: 10.15439/2015F393
ACSIS, Vol. 6

Consistency-Based Preprocessing for Classification of Data Coming from Evaluation Sheets of Subjects with ASDs

Krzysztof Pancerz, Aneta Derkacz
University of Management and Administration
Zamość, Poland
Email: kpancerz@wszia.edu.pl

Jerzy Gomuła
Cardinal Stefan Wyszyński University
Warsaw, Poland
Email: jerzy.gomula@wp.pl

***Abstract*—In general, the aim of our research is to adapt computational intelligence methods for computer-aided decision support in diagnosis and therapy of persons with Autism Spectrum Disorders (ASDs). In the paper, we are focusing on the data preprocessing step for cleaning a training data set for classifiers. An approach based on consistency factors is proposed.**

I. Introduction

AUTISM is a brain development disorder that impairs social interaction and communication, and causes restricted and repetitive behaviors, all starting before a child is three years old. Starting in May 2013, i.e., the date of publication of the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5), all autism disorders were merged into one umbrella diagnosis of Autism Spectrum Disorders (ASDs). Autism spectrum disorders can dramatically affect a child's life, as well as that of their families, schools, friends and a wider community. Therefore, we decided to start research on adaptation of computational intelligence methods, with particular regard to data mining and machine learning ones, for computer aided-decision support in diagnosis and therapy of persons with autism spectrum disorders (ASDs). Computer-based decision support (CDS) is defined as the use of a computer to bring relevant knowledge to bear on the health care and well-being of a patient [1]. Input data come from original author's evaluation sheets of subjects with ASDs in the important spheres (among others, self-service, communication, cognitive, physical, as well as the sphere responsible for functioning in the social and family environment, etc.). Computer-aided analysis enables us to determine trends in the abovementioned spheres (progress, stagnation, or regress) and support adjustments of the individual therapeutic and educational programs for persons covered by the care.

II. Input Data

Experiments testing the relative effectiveness of our approach have been performed on data describing over 70 cases (subjects) classified into three categories: high-functioning, medium-functioning, or low-functioning autism. Each subject has been evaluated using an original author's sheet including questions about competencies grouped into 17 spheres marked with Roman numerals:

- VI. Support for active communication.
- VII. Active communication concerning objects, people, parts of the body.
- VIII. Imitation, the length and complexity of the utterance.
- IX. Needs, emotions, moods.
- X. Object communication (the level of specific symbols).
- XI. Symbolic communication.
- XII. Requests.
- XIII. Choices.
- XIV. Communication in a pair (with contemporary, with an adult).
- XV. Social communication competences.
- XVI. Communication in a group and in social situations (in a team, in school, in the closest social environment).
- XVIII. Vocabulary.
- XIX. The degree of effectiveness of information.
- XX. The degree of motivation to communicate.
- XXI. The degree and type of hint in communication.
- XXII. Building the utterance - the degree of its complexity and functionality.
- XXIII. Dialogues.

Each case x is described by a data vector $a(x)$ consisting of over 300 descriptive attributes: $a(x) = [a_1(x), a_2(x), ..., a_m(x)]$. Such a data vector is called a profile. Four values of descriptive attributes are possible, namely 0, 25, 50, and 100. They have the following meaning:

- 0 - not performed,
- 25 - performed after physical help,
- 50 - performed after verbal help/demonstration,
- 100 - performed unaided.

If we have training data for classifiers, then to each case x we also add one decision attribute c - a class (category) to which a patient is classified. For decision attribute values, we use the following notation:

- LOW - low-functioning autism,
- $MEDIUM$ - medium-functioning autism,
- $HIGH$ - high-functioning autism.

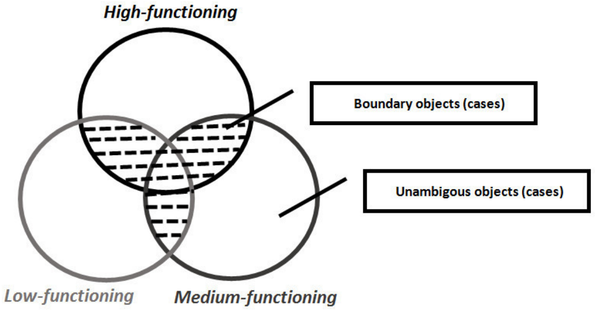


Fig. 1. Dividing a set of all training objects (cases)

In the current stage of research, each sphere is treated separately. For each sphere, the training data (which are used to learn or extract relationships between data) are stored in a tabular form (see example in Table I) which is formally called a decision table.

A decision table represents a decision system in the Pawlak's form (cf. [2]). We use the following formal definition of a decision system. A decision system DS is a tuple $DS = (U, C, D, V_{con}, V_{dec}, f_{inf}, f_{dec})$, where:

- U is a nonempty, finite set of objects,
- C is a nonempty, finite set of condition attributes,
- D is a nonempty, finite set of decision attributes,
- $V_{con} = \bigcup_{c \in C} V_c$, where V_c is a set of values of the condition attribute c,
- $V_{dec} = \bigcup_{d \in D} V_d$, where V_d is a set of values of the decision attribute d,
- $f_{inf} : C \times U \to V_{con}$ is an information function such that $f_{inf}(c, u) \in V_c$ for each $c \in C$ and $u \in U$,
- $f_{dec} : D \times U \to V_{dec}$ is a decision function such that $f_{dec}(d, u) \in V_d$ for each $d \in D$ and $u \in U$.

III. Preprocessing

Preprocessing is an important stage in data mining and knowledge discovery processes. It encompasses different tasks, e.g., extraction and selection of attributes (features), discretization of attribute values, data cleaning, etc. In this section, we describe some kind of data cleaning which is used as a preprocessing step in classification of data coming from evaluation sheets of subjects with ASDs. In our approach to classification, we can distinguish the following main stages:

1) Calculating consistency factors of objects included in the decision subsystem corresponding to class Y, with the knowledge included in the decision subsystem corresponding to class X.
2) Dividing a set of all training objects (cases) into two subsets:
   - a subset of unambiguous objects (cases),
   - a subset of boundary objects (cases).
3) Building separate classifiers trained on unambiguous objects and boundary objects, respectively.

The main aim of Stage 1 is to determine two subsets of objects included in a training data set: a subset of unambiguous objects (cases) as well as a subset of boundary objects (cases), see Figure 1.

Let $DS = (U, C, D, V_{con}, V_{dec}, f_{inf}, f_{dec})$ be a decision system, where $D = \{d\}$ and $V_{dec} = \{v_{d_1}, v_{d_2}, \ldots, v_{d_k}\}$. The set U of objects can be divided into disjoint subsets according to values of a decision attribute d, i.e.:

$$\bigcup_{i=1,2,\ldots,k} X_i,$$

where:

- $X_1 \cap X_2 \cap \cdots \cap X_k = \emptyset$,
- $X_1 \cup X_2 \cup \cdots \cup X_k = U$.

An object $u \in U$ is called a boundary object if it belongs to the subset X_i, where $i = 1, 2, \ldots, k$ and there exists X_j, where $j = 1, 2, \ldots, k$ and $j \neq i$ such that the consistency factor of u with the knowledge included in X_j is greater or equal to a given threshold θ, where $\theta \in [0, 1]$.

To differentiate two subsets of objects (unambiguous objects and boundary objects), we use an approach based on consistency factors. We assume that the boundary objects should be treated individually in a process of training the classifier (see Figure 2) because they are assigned to one decision class but they are also closed to other decision classes with respect to consistency factors. Boundary objects are intended for training more specialized and sensitive classifiers.

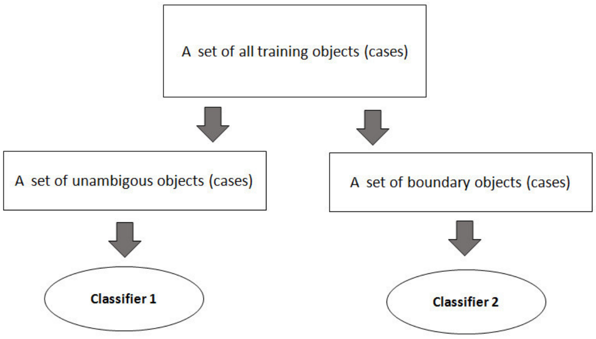

A set of all training objects (cases)
A set of unambiguous objects (cases)
A set of boundary objects (cases)
Classifier 1
Classifier 2

Fig. 2. Building separate classifiers

A decision system includes a finite set of cases described by attributes. Each attribute represents one of the features of cases. Apart from all cases included in the original decision system, we can consider some other cases. Such cases can be totally consistent or consistent to a certain degree with the knowledge included in the original system. The knowledge can be represented in the form of rules (production, association, etc.), cf. [3], [4]. The problem is to determine consistency factors of new cases taken into consideration with the knowledge included in the original decision system. We have adopted calculation of the consistency factor according to the definition used in [5]. That definition was derived from the approach to computing consistency factors of objects in information systems proposed in [4]. It is worth noting that an information system differs from a decision system only by the lack of decision attributes. A formal definition is as follows.

TABLE I
EXEMPLARY INPUT DATA COMING FROM THE EVALUATION SHEET

ID	VI.117	...	VI.120	VI.120a	...	VI.120f	VI.121a	...	VI.121g	VI.122	$class$
#1	50	...	100	50	...	0	50	...	0	0	LOW
...	...	...	...	...	...	...	...	...	...	...	...
#32	25	...	100	100	...	25	50	...	50	0	$MEDIUM$
...	...	...	...	...	...	...	...	...	...	...	...
#66	0	...	0	0	...	0	100	...	100	100	$HIGH$
...	...	...	...	...	...	...	...	...	...	...	...

An information system IS is a quadruple $IS = (U, A, V, f)$, where:

- U is a nonempty, finite set of objects,
- A is a nonempty, finite set of attributes,
- $V = \bigcup\limits_{a \in A} V_a$, where V_a is a set of values of the attribute a,
- $f : A \times U \rightarrow V$ is an information function such that $f(a, u) \in V_a$ for each $a \in A$ and $u \in U$.

It is assumed, in the algorithm for computing a consistency factor, that the knowledge included in an original information system S is expressed by minimal rules true and realizable in S. Computing a consistency factor for a given object is based on determining importance (relevance) of rules extracted from the system S which are not satisfied by the new case. If the importance of these rules is greater the consistency factor of a new object with the knowledge is smaller. The importance of a set of rules not satisfied by the new case is determined by means of a strength factor of this set of rules in S. This approach has been implemented in CLAPSS (Classification and Prediction Software System) - a computer tool for solving different classification and prediction problems using, among others, some specialized approaches based mainly on the rough set theory (see [6]). The tool was designed for the Java platform. The main features of CLAPSS are the following:

- Portability. Thanks to the Java technology, the application works on various software and hardware platforms. In the future, the tool can be adapted for platforms available in mobile devices and as a service in the cloud.
- User-friendly interface (see Figure 3).
- Modularity. The project of CLAPSS and its implementation takes into consideration modularity. It makes CLAPSS possible to easily extend in the future.

Consistency factors are calculated in CLAPSS using the algorithm based on rough sets given in [7]. This algorithm makes use of important results of research on extensions of information systems given in [8]. Therefore, we recall crucial notions concerning rough sets. For more exact description and explanation we refer readers to [2] and [9].

Let $IS = (U, A, V, f)$ be an information system. Each subset $B \subseteq A$ of attributes determines an equivalence relation on U, called an indiscernibility relation $Ind(B)$, defined as

$$Ind(B) = \{(u, v) \in U \times U : \forall_{a \in B} f(a, u) = f(a, v)\}.$$

The equivalence class containing $u \in U$ will be denoted by $[u]_B$.

Let $X \subseteq U$ and $B \subseteq A$. The B-lower approximation $\underline{B}X$ of X and the B-upper approximation $\overline{B}X$ of X are defined as

$$\underline{B}X = \{u \in U : [u]_B \subseteq X\}$$

and

$$\overline{B}X = \{u \in U : [u]_B \cap X \neq \emptyset\},$$

respectively. A set $BN_B(X) = \overline{B}(X) - \underline{B}(X)$ is called the B-boundary region of X. The B-lower approximation $\underline{B}X$ of X is the set of all objects from U, which can be for certain classified as X using B, i.e., they are certainly X in view of B. The B-upper approximation $\overline{B}X$ of X is the set of all objects from U, which can be possibly classified as X using B, i.e., they are possibly X in view of B. The B-boundary region $BN_B(X)$ of X is the set of all objects from U, which can be classified neither as X nor as not-X using B. If $BN_B(X) = \emptyset$, then X is sharp (exact) with respect to B. Otherwise, X is rough (inexact).

We can provide the definition of a consistency factor (cf. [7] and [5]) in terms of appropriate lower approximations of sets. Let

- $A_{\widetilde{a}} = A - \{a\}$, where $a \in A$,
- $X_a^v = \{u \in U : f(a, u) = v\}$,
- and $\widetilde{U} = \bigcup\limits_{a \in A} \bigcup\limits_{v \in V_a} \{\underline{A_{\widetilde{a}}}(X_a^v) : \underline{A_{\widetilde{a}}}(X_a^v) \neq \emptyset \wedge f(a, u^*) \neq v\}$.

The consistency factor $\xi_{IS}(u^*)$ of u^* is defined as follows:

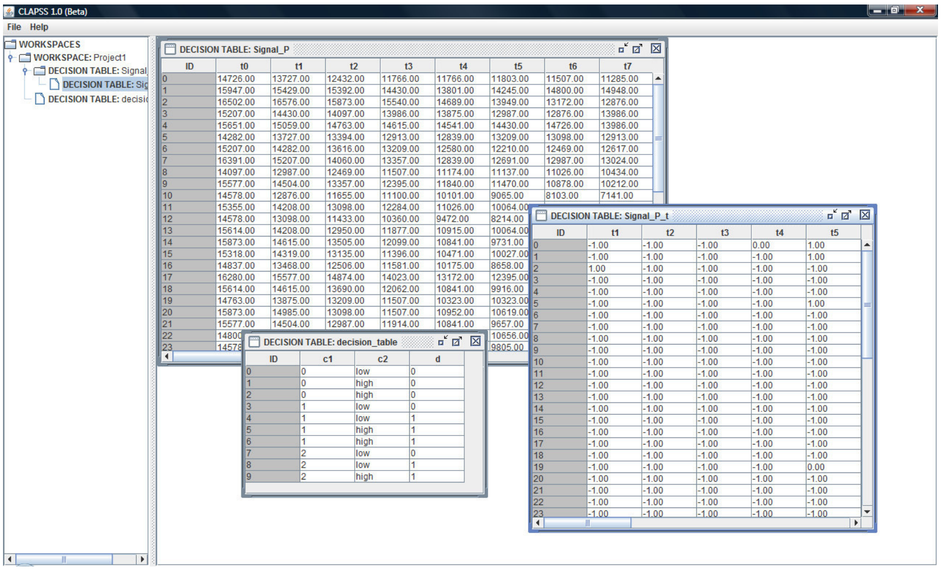
$$\xi_{IS}(u^*) = \xi'_{IS}(u^*)\omega_{IS}(u^*),$$

where:

- $\xi'_{IS}(u^*) = 1 - \frac{card(\widetilde{U})}{card(U)}$ is a proper consistency,
- $\omega_{IS}(u^*) = \frac{card(\{a \in A : f(a, u^*) \in V_a\})}{card(A)}$ is a resemblance factor determining some affinity between the object u and objects from IS with respect to values of attributes.

A general scheme of calculating consistency factors for determining unambiguous and boundary objects is shown in Figure 4.

In experiments, for subsets of unambigous objects (cases), we have noticed significant improvement of classification accuracy (sometimes more than 10 percentage points).



DECISION TABLE: Signal_P

ID	t0	t1	t2	t3	t4	t5	t6	t7
0	14726.00	13727.00	12432.00	11766.00	11766.00	11803.00	11507.00	11285.00
1	15947.00	15429.00	15392.00	14430.00	13801.00	14245.00	14800.00	14948.00
2	16502.00	16576.00	15873.00	15540.00	14689.00	13949.00	13172.00	12876.00
3	15207.00	14430.00	14097.00	13986.00	13875.00	12987.00	12876.00	13986.00
4	15651.00	15059.00	14763.00	14615.00	14541.00	14430.00	14726.00	13986.00
5	14282.00	13727.00	13394.00	12913.00	12839.00	13209.00	13098.00	12913.00
6	15207.00	14282.00	13616.00	13209.00	12580.00	12210.00	12469.00	12617.00
7	16391.00	15207.00	14060.00	13357.00	12839.00	12691.00	12987.00	13024.00
8	14097.00	12987.00	12469.00	11507.00	11174.00	11137.00	11026.00	10434.00
9	15577.00	14504.00	13357.00	12395.00	11840.00	11470.00	10878.00	10212.00
10	14578.00	12876.00	11655.00	11100.00	10101.00	9065.00	8103.00	7141.00
11	15355.00	14208.00	13098.00	12284.00	11026.00	10064.00		
12	14578.00	13098.00	11433.00	10360.00	9472.00	8214.00		
13	15614.00	14208.00	12950.00	11877.00	10915.00	10064.00		
14	15873.00	14615.00	13505.00	12099.00	10841.00	9731.00		
15	15318.00	14319.00	13135.00	11396.00	10471.00	10027.00		
16	14837.00	13468.00	12506.00	11581.00	10175.00	8658.00		
17	16280.00	15577.00	14874.00	14023.00	13172.00	12395.00		
18	15614.00	14615.00	13690.00	12062.00	10841.00	9916.00		
19	14763.00	13875.00	13209.00	11507.00	10323.00	10323.00		
20	15873.00	14985.00	13098.00	11507.00	10952.00	10619.00		
21	15577.00	14504.00	12987.00	11914.00	10841.00	9657.00		
22	14800					10656.00		
23	14578					9805.00		

DECISION TABLE: Signal_P_t

ID	t1	t2	t3	t4	t5
0	-1.00	-1.00	-1.00	0.00	1.00
1	-1.00	-1.00	-1.00	-1.00	1.00
2	1.00	-1.00	-1.00	-1.00	-1.00
3	-1.00	-1.00	-1.00	-1.00	-1.00
4	-1.00	-1.00	-1.00	-1.00	-1.00
5	-1.00	-1.00	-1.00	-1.00	1.00
6	-1.00	-1.00	-1.00	-1.00	-1.00
7	-1.00	-1.00	-1.00	-1.00	-1.00
8	-1.00	-1.00	-1.00	-1.00	-1.00
9	-1.00	-1.00	-1.00	-1.00	-1.00
10	-1.00	-1.00	-1.00	-1.00	-1.00
11	-1.00	-1.00	-1.00	-1.00	-1.00
12	-1.00	-1.00	-1.00	-1.00	-1.00
13	-1.00	-1.00	-1.00	-1.00	-1.00
14	-1.00	-1.00	-1.00	-1.00	-1.00
15	-1.00	-1.00	-1.00	-1.00	-1.00
16	-1.00	-1.00	-1.00	-1.00	-1.00
17	-1.00	-1.00	-1.00	-1.00	-1.00
18	-1.00	-1.00	-1.00	-1.00	-1.00
19	-1.00	-1.00	-1.00	-1.00	0.00
20	-1.00	-1.00	-1.00	-1.00	-1.00
21	-1.00	-1.00	-1.00	-1.00	-1.00
22	-1.00	-1.00	-1.00	-1.00	-1.00
23	-1.00	-1.00	-1.00	-1.00	-1.00

DECISION TABLE: decision_table

ID	c1	c2	d
0	0	low	0
1	0	high	0
2	0	high	0
3	1	low	0
4	1	low	1
5	1	high	1
6	1	high	1
7	2	low	0
8	2	low	1
9	2	high	1

Fig. 3. User-friendly interface of CLAPSS

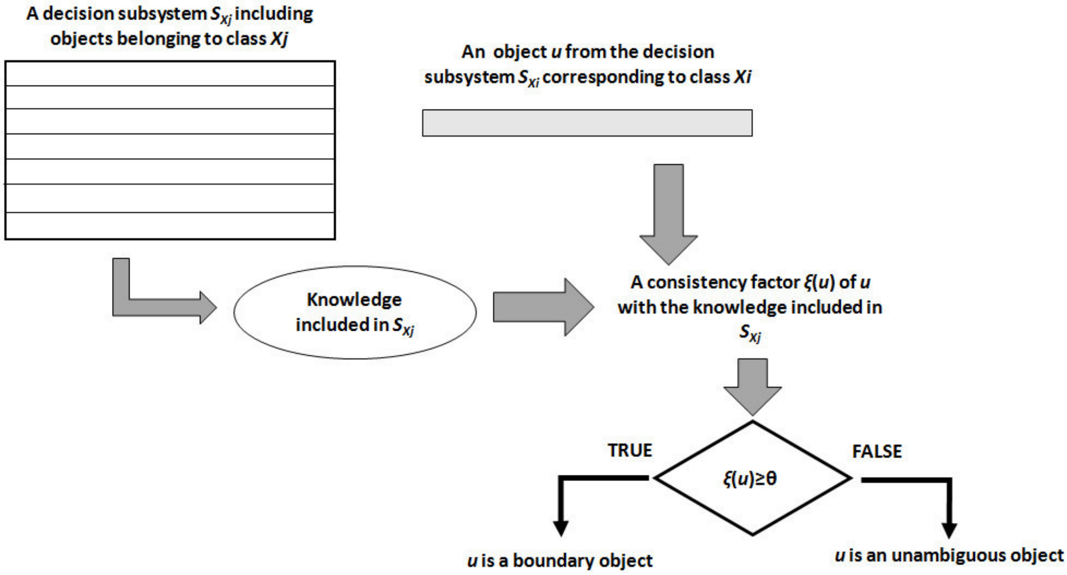

A decision subsystem S_{Xj} including objects belonging to class *Xj*
An object *u* from the decision subsystem S_{Xi} corresponding to class *Xi*
Knowledge included in S_{Xj}
A consistency factor ξ(*u*) of *u* with the knowledge included in S_{Xj}
ξ(*u*)≥θ
TRUE
FALSE
u is a boundary object
u is an unambiguous object

Fig. 4. Calculating consistency factors for determining unambiguous and boundary objects

IV. Conclusions and Further Work

We have described initial research on computer-aided analysis of data coming from evaluation sheets of subjects with autism spectrum disorders. This stage of research is focused on the data preprocessing step. An approach to clean a training data set for classifiers, based on consistency factors, has been proposed. The important problem in the future is to determine consistency factors of new cases taking into consideration different ways of knowledge representation. In the further stages of research, we will be interested in building hybrid classifiers combining a wide range of approaches. Adopted methods will be implemented in the specialized computer tool modelled on our previous tool, called Copernicus [10], intended for analysis and classification of data coming from the MMPI (Minnesota Multiphasic Personality Inventory) test (cf. [11]).

References

[1] R. Greenes, *Clinical Decision Support. The Road Ahead*. Elsevier Inc., 2007.

[2] Z. Pawlak, *Rough Sets. Theoretical Aspects of Reasoning about Data*. Dordrecht: Kluwer Academic Publishers, 1991.

[3] Z. Suraj, “Some remarks on extensions and restrictions of information systems,” in *Rough Sets and Current Trends in Computing*, ser. Lecture Notes in Artificial Intelligence, W. Ziarko and Y. Yao, Eds. Berlin Heidelberg: Springer Verlag, 2001, vol. 2005, pp. 204–211.

[4] Z. Suraj, K. Pancerz, and G. Owsiany, "On consistent and partially consistent extensions of information systems," in *Proceedings of the RSFDGrC'2005*, ser. Lecture Notes in Artificial Intelligence, D. Ślęzak *et al.*, Eds. Berlin Heidelberg: Springer Verlag, 2005, vol. 3641, pp. 224–233.

[5] Ł. Piątek, K. Pancerz, and G. Owsiany, "Validation of data categorization using extensions of information systems: Experiments on melanocytic skin lesion data," in *Proceedings of the FedCSIS'2011*, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., Szczecin, Poland, 2011, pp. 147–151.

[6] K. Pancerz, "On selected functionality of the classification and prediction software system (CLAPSS)," in *Proceedings of the IDT'2015*, Zilina, Slovakia, 2015, pp. 267–274.

[7] ——, "Extensions of information systems: The rough set perspective," ser. Lecture Notes in Computer Science, J. Peters, A. Skowron, M. Chakraborty, W.-Z. Wu, and M. Wolski, Eds. Berlin Heidelberg: Springer-Verlag, 2009, vol. 5656, pp. 157–168.

[8] M. Moshkov, A. Skowron, and Z. Suraj, "On testing membership to maximal consistent extensions of information systems," in *Rough Sets and Current Trends in Computing*, ser. Lecture Notes in Artificial Intelligence, S. Greco, Y. Hata, S. Hirano, M. Inuiguchi, S. Miyamoto, H. S. Nguyen, and R. Slowinski, Eds. Berlin Heidelberg: Springer-Verlag, 2006, vol. 4259, pp. 85–90.

[9] Z. Pawlak and A. Skowron, "Rudiments of rough sets," *Information Sciences*, vol. 177, pp. 3–27, 2007. doi: 10.1016/j.ins.2006.06.003

[10] K. Pancerz, O. Mich, A. Burda, and J. Gomuła, "A tool for computer-aided diagnosis of psychological disorders based on the mmpi test: an overview," in *Applications of Computational Intelligence in Biomedical Technology*, ser. Studies in Computational Intelligence, R. Bris, J. Majernik, K. Pancerz, and E. Zaitseva, Eds. Springer International Publishing, 2016, vol. 606, pp. 201–213.

[11] D. Lachar, *The MMPI: Clinical assessment and automated interpretations*. Fate Angeles: Western Psychological Services, 1974.