Rule-Based Classification of Patients Screened with the MMPI Test in the Copernicus System

Daniel Jachyra, Jerzy Gomuła and Krzysztof Pancerz

Abstract The Copernicus system is a tool for computer-aided diagnosis of mental disorders based on personality inventories. Knowledge representation in the form of rules is the closest method to human activity and reasoning, among others, in making a medical diagnosis. Therefore, in the Copernicus system, rule-based classification of patients screened with the MMPI test is one of the most important parts of the tool. The main goal of the chapter is to give more precise view of this part of the developed tool.

1 Introduction

Computer support systems supporting medical diagnosis have become increasingly more popular worldwide (cf. [14]). Therefore, the tool called Copernicus [11] for classification of patients with mental disorders screened with the MMPI test has been developed. The Minnesota Multiphasic Personality Inventory (MMPI) test (cf. [6, 15]) delivering psychometric data on patients with selected mental

D. Jachyra (✉)
Chair of Information Systems Applications, University of Information Technology
and Management in Rzeszów, Rzeszów, Poland
e-mail: djachyra@wsiz.rzeszow.pl

J. Gomuła
The Andropause Institute, Medan Foundation, Warsaw, Poland
e-mail: jerzy.gomula@wp.pl

J. Gomuła
Cardinal Stefan Wyszyński University in Warsaw, Warsaw, Poland

K. Pancerz
Institute of Biomedical Informatics, University of Information Technology
and Management in Rzeszów, Rzeszów, Poland
e-mail: kpancerz@wsiz.rzeszow.pl

disorders is one of the most frequently used personality tests in clinical mental health as well as psychopathology (mental and behavioral disorders). In years 1998–1999, a team of researchers, consisting of W. Duch, T. Kucharski, J. Gomuła, R. Adamczak, created two independent rule systems, devised for the nosological diagnosis of persons, that may be screened with the MMPI-WISKAD test [7]. The MMPI-WISKAD personality inventory is a Polish adaptation of the American inventory (see [3, 16]). The Copernicus system developed by us is the continuation and expansion of that research.

In the Copernicus system, different quantitative groups of methods supporting differential inter-profile diagnosis have been selected and implemented. However, rule-based classification of patients is one of the most important parts of the tool. Knowledge representation in the form of rules is the closest method to human activity and reasoning, among others, in making a medical diagnosis. In the most generic format, medical diagnosis rules are conditional statements of the form:

$$\text{IF } conditions(symptoms), \text{ THEN } decision(diagnosis).$$

The rule expresses the relationship between symptoms determined on the basis of examination and diagnosis which should be taken for these symptoms before the treatment. In our case, symptoms are determined on the basis of results of a patient's examination using the MMPI test. In the Copernicus system, the number of rule sets generated by different data mining and machine learning algorithms (for example: the RSES system [1], the WEKA system [17]) is included. However, the rule base can be extended to new rule sets delivered by the user. In the case of multiple sources of rules, we obtain a combination of classifiers. In classifier combining, predictions of classifiers should be aggregated into a single prediction in order to improve the classification quality. The Copernicus system delivers a number of aggregation functions described in the remaining part of this chapter.

The Copernicus system supports the idea that visualization plays an important role in professional decision support. Some pictures often represent data better than expressions or numbers. Visualization is very important in dedicated and specialized software tools used in different (e.g., medical) communities. In the Copernicus system, a special attention has been paid to the visualization of analysis of MMPI data for making a diagnosis decision easier. A unique visualization of classification rules in the form of stripes put on profiles as well as visualization of results of aggregated classification have been designed and implemented.

2 MMPI Data

In the case of the MMPI test, each case (patient) x is described by a data vector $a(x)$ consisting of thirteen descriptive attributes: $a(x) = [a_1(x), a_2(x), \ldots, a_{13}(x)]$. If we have training data, then to each case x we additionally add one decision attribute d—a class to which a patient is classified.

The validity part of the profile consists of three scales: L (laying), F (atypical and deviational answers), K (self defensive mechanisms). The clinical part of the profile consists of ten scales: 1.Hp (Hypochondriasis), 2.D (Depression), 3.Hy (Hysteria), 4.Ps (Psychopathic Deviate), 5.Mf (Masculinity/Femininity), 6.Pa (Paranoia), 7.Pt (Psychasthenia), 8.Sc (Schizophrenia), 9.Ma (Hypomania), 0.It (Social introversion). The clinical scales have numbers attributed so that a profile can be encoded to avoid negative connotations connected with the names of scales. Values of attributes are expressed by the so-called T-scores. The T-scores scale, which is traditionally attributed to MMPI, represents the following parameters: offset ranging from 0 to 100 T-scores, average equal to 50 T-scores, standard deviation equal to 10 T-scores.

For our research, we have obtained input data which has nineteen nosological classes and the reference class (*norm*) assigned to patients by specialists. Each class corresponds to one of psychiatric nosological types: neurosis (*neur*), psychopathy (*psych*), organic (*org*), schizophrenia (*schiz*), delusion syndrome (*del.s*), reactive psychosis (*re.psy*), paranoia (*paran*), sub-manic state (*man.st*), criminality (*crim*), alcoholism (*alcoh*), drug addiction (*drug*), simulation (*simu*), dissimulation (*dissimu*), and six deviational answering styles (*dev1, dev2, dev3, dev4, dev5, dev6*). The data set examined in the Copernicus system was collected by T. Kucharski and J. Gomuła from the Psychological Outpatient Clinic.

Data vectors can be represented in a graphical form as the so-called MMPI profiles. The profile always has a fixed and invariable order of its constituents (attributes, scales). Let a patient x be described by the data vector:

$$a(x) = [56, 78, 55, 60, 59, 54, 67, 52, 77, 56, 60, 68, 63].$$

Its profile is shown in Fig. 1.

A basic profile can be extended by additional indexes or systems of indexes (cf. [13]). Different combinations of scales constitute diagnostically important indexes (e.g., Gough's, Goldberg's, Watson-Thomas's, L'Abate's, Lovell's indexes—see [6]) and systems of indexes (e.g., Diamond's [5], Leary's, Eichmann's, Petersen's, Taulbee-Sisson's, Butcher's [2], Pancheri's). They have been determined on the basis of clinical and statistical analysis of many patients' profiles. All mentioned and some additional (e.g., the nosological difference-configuration Gough-Płużek's system) indexes have been implemented in the Copernicus system. It enables the user to extend the basic profile even to 100 attributes (13 scales plus 87 indexes).

Table 1 An input data for Copernicus (fragment)

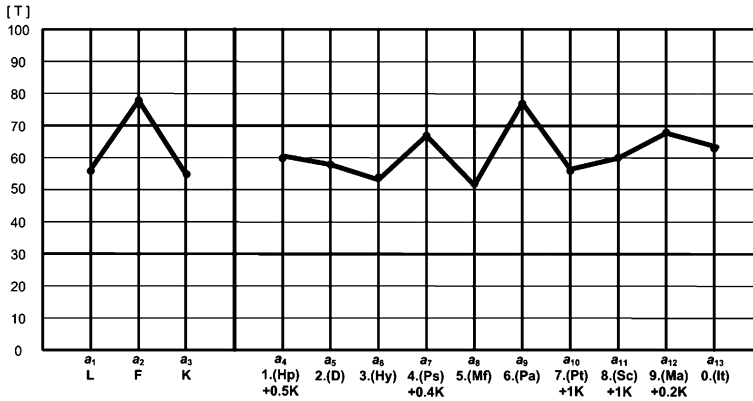| Attribute Scale	a_1 L	a_2 F	a_3 K	a_4 1.Hp	a_5 2.D	a_6 3.Hy	a_7 4.Ps	a_8 5.Mf	a_9 6.Pa	a_{10} 7.Pt	a_{11} 8.Sc	a_{12} 9.Ma	a_{13} 0.It	class
#1	55	65	50	52	65	57	63	56	61	61	60	51	59	*norm*
#2	50	73	53	56	73	63	53	61	53	60	69	45	61	*org*
#3	56	78	55	60	59	54	67	52	77	56	60	68	63	*paran*
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Fig. 1 MMPI profile of a patient (example); suppressors +0.5K, +0.4K, +1K, +0.2K—a correction value from raw results of scale K added to raw results of selected clinical scales

3 Rule-Based Classification in Copernicus

In this section, we describe step-by-step functionality of the Copernicus system concerning rule-based classification of patients screened with the MMPI test.

3.1 General Selection of Rule Sets

For classification purposes, the user can select a number of rule sets included in the tool (see Fig. 2). Such rule sets have been generated by different data mining and machine learning algorithms (for example: the RSES system [1], the WEKA system [17]). Moreover, the user can select its own rule set. The problem of selecting suitable sets of rules for classification of MMPI profiles has been considered in our previous chapters (see [8–10, 12, 13]).

3.2 Specific Selection of Rules

After general selection of rule sets, the user can determine more precisely which rules will be used in the classification process.

Each rule R in the Copernicus system has the form:

$$\text{IF } a_{i1}(x) \in [x_{i1}^l, x_{i1}^r] \text{ AND } \ldots \text{ AND } a_{ik}(x) \in [x_{ik}^l, x_{ik}^r], \text{ THEN } d(x) = d_m, \quad (1)$$

where $a_{i1}, \ldots, a_{ik}$ are selected scales (validity and clinical) or specialized indexes, $x_{i1}^l, x_{i1}^r, \ldots, x_{ik}^l, x_{ik}^r$ are the left and right endpoints of intervals, respectively, d is a
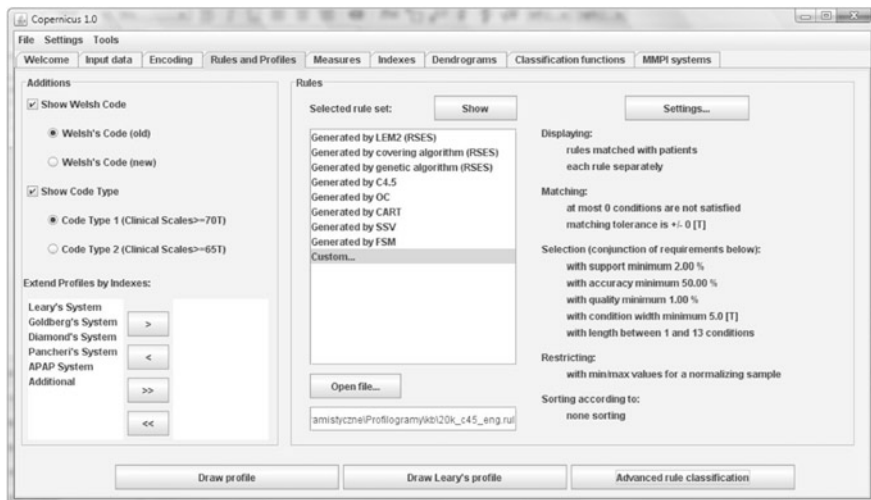
Fig. 2 The rule selection window in the Copernicus system

diagnosis, d_m is one of nosological classes proposed for the diagnosis. Each $a_{ik}(x) \in [x_{ik}^l, x_{ik}^r]$ is called an elementary condition of R. For each patient x, its profile is said to be matched to a rule R if and only if $a_{i1}(x) \in [x_{i1}^l, x_{i1}^r]$ and ... and $a_{ik}(x) \in [x_{ik}^l, x_{ik}^r]$. This fact is denoted by $x| = R$.

Each rule set included in the Copernicus system has been extracted from a proper set of cases called a training set. Each rule R in the form of 3.1 can be characterized by the following factors:

- the accuracy factor $acc(R) = \frac{n_C}{n_{CD}}$,
- the total support factor $supp_t(R) = \frac{n_C}{n}$,
- the class support factor $supp_c(R) = \frac{n_C}{n_D}$,
- the quality factor $qual(R) = acc(R)supp(R)$,
- the length factor $length(R) = card(a_{i1}, ..., a_{ik})$, i.e., a number of elementary conditions of R,

where *card* denotes the cardinality of a given set, n is the size (a number of cases) of the training set, n_C is a number of cases in the training set which are matched to R, n_{CD} is a number of cases in the training set which are matched to R and which have additionally assigned the class d_m, n_D is a number of cases in the training set which have assigned the class d_m.

Exemplary classification rules obtained from data consisting of profiles using the WEKA system are shown in Table 2. The accuracy factor has been assigned to each rule.

The user can set, among others, that each rule R used for classification a given case x satisfies the following conditions:

Table 2 Exemplary rules obtained after transformation of a decision tree generated for all scales excluding scale 5

L	1.*Hp*	3.*Hy*	6.*Pa*	8.*Sc*	9.*Ma*	0.*It*	Rule No.:class (accuracy %)
<=58	<=57						R1:*norm* (89)
	58 − 64	<=59	<=77	<=68			R2:*norm* (89
	>64		<=77		<=56		R3:*neur* (85)
	58 − 64	>59	<=77	<=68			R4:*neur* (85)
>58	<=57		>58				R5:*psych* (92)
	>64		<=77		>56		R6:*schiz* (85)
	58 − 64		<=77	>68			R7:*schiz* (85)
>58	<=57		<=58			>59	R8:*simu* (94)
	>57		>77				R9:*simu* (94)
>58	<=57		<=58			<=59	R10:*dissimu*(85)

- a number of elementary conditions of R for which x does not match R is equal to 0 (exact matching) or more (approximate matching),
- x is matched to R with a certain degree (tolerance), i.e., for each elementary condition $a_{ik}(x) \in [x_{ik}^l - t, x_{ik}^r + t]$, where t is a tolerance value,
- $acc(R)$ is greater or equal to a given threshold,
- $supp(R)$ is greater or equal to a given threshold,
- $qual(R)$ is greater or equal to a given threshold,
- the interval $[x_{ik}^l, x_{ik}^r]$ of each elementary condition of R has the property that $x_{ik}^r - x_{ik}^r$ is greater or equal to a given threshold (i.e., a rule with tight conditions can be omitted),
- $length(R)$ is included in a given interval (i.e., a rule with too short or too long condition part can be omitted),
- indicated scales in elementary conditions of R are omitted—the so called scale excluding (for example diagnosticians' experience shows that the scale 5.*Mf* is weak and it should be omitted).

For elementary conditions in the form of intervals, we sometimes obtain lower and upper bounds that are, for example, $-\infty$ and ∞, respectively. Such values cannot be rationally interpreted from the clinical point of view. Therefore, ranges of classification rule conditions can be restricted. We can replace ∞ by:

- a maximal value of a given scale occurring for a given class in our sample,
- a maximal value of a given scale for all twenty classes,
- a maximal value of a given scale for a normalizing group (i.e., a group of women, for which norms of validity and clinical scales have been determined),
- a maximal value for all scales for a normalizing group, i.e., 120 T-scores.

A procedure for restricting ranges of classification rule conditions with the value $-\infty$ is carried out similarly, but we take into consideration minimal values. A minimal value for all scales of normalizing group of women is 28 T-scores. Specialized indexes are linear combinations of scales. Therefore, restricting ranges

of rule conditions for indexes is also possible and simple. Copernicus enables us to select the way of restricting ranges. Rule conditions are automatically restricted to the form readable for the diagnostician-clinician.

3.3 Visualization of Profiles and Rules

The rule R can be graphically presented as a set of stripes placed in the profile space. Each condition part $a_{ij}(x) \in [x_{ij}^l, x_{ij}^r]$, where $j = 1, ..., k$, of the rule R is represented as a vertical stripe on the line corresponding to the scale a_{ij}. This stripe is restricted from both the bottom and the top by values x_{ij}^l, x_{ij}^r, respectively. Such visualization enables the user to easily determine which rule matches a given profile (cf. Fig. 3).

3.4 Classification Results

On the basis of rules a proper diagnostic decision for the case x can be made. Aggregation factors implemented in the Copernicus system enable selecting only one main decision from decisions provided by rules used for the classification of x. For each class d for which cases can be classified, a number of different aggregation factors can be calculated. The first aggregation factor is the simplest one.
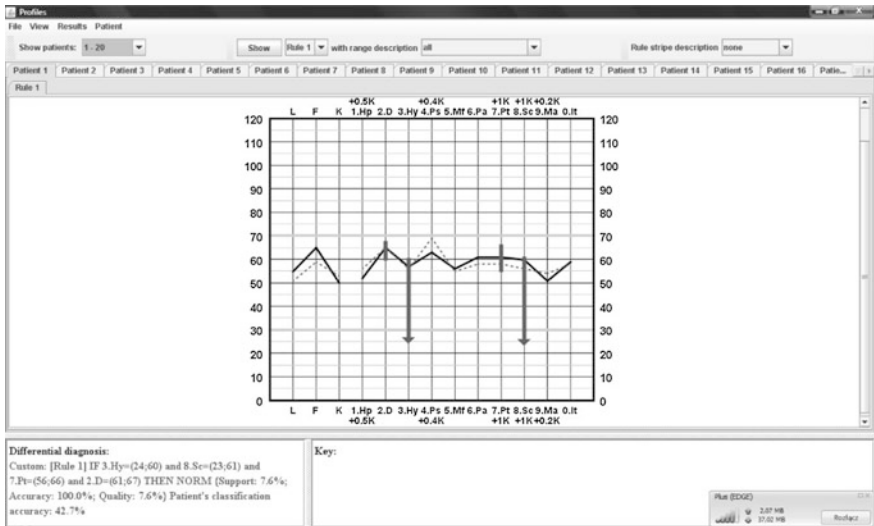
Fig. 3 Visualization of a rule in the profile space

It expresses the relative number of rules denoting the class d in the set of all rules matched by x in the selected sense (see Sect. 3.2):

$$aggr_1(d) = \frac{card(\{R : x| = R \ and \ class(R) = d\})}{card(\{R : x| = R\})}.$$

Another three aggregation factors take also into consideration the maximal value of quality factors of rules from indicating the class d. These factors differ on the weights of components:

$$aggr_2(d) = 0.8 \max(\{qual(R) : x| = R \ and \ class(R) = d\})$$
$$+ 0.2 \frac{card(\{R : x| = R \ and \ class(R) = d\})}{card(\{R : x| = R\})},$$

$$aggr_3(d) = 0.5 \max(\{qual(R) : x| = R \ and \ class(R) = d\})$$
$$+ 0.5 \frac{card(\{R : x| = R \ and \ class(R) = d\})}{card(\{R : x| = R\})},$$

$$aggr_4(d) = 0.67 \max(\{qual(R) : x| = R \ and \ class(R) = d\})$$
$$+ 0.33 \frac{card(\{R : x| = R \ and \ class(R) = d\})}{card(\{R : x| = R\})}.$$

The last two aggregation factors additionally take into consideration the average length of rules from the d class. In this case, the smaller the average length is, the better the set of rules. These factors differ on the weights of components:

$$aggr_5(d) = 0.6 \max(\{qual(R) : x| = R \ and \ class(R) = d\})$$
$$+ 0.2 \text{avg}(\{1 - length(R) : x| = R \ and \ class(R) = d\})$$
$$+ 0.2 \frac{card(\{R : x| = R \ and \ class(R) = d\})}{card(\{R : x| = R\})},$$

$$aggr_6(d) = 0.4 \max(\{qual(R) : x| = R \ and \ class(R) = d\})$$
$$+ 0.4 \text{avg}(\{1 - length(R) : x| = R \ and \ class(R) = d\})$$
$$+ 0.2 \frac{card(\{R : x| = R \ and \ class(R) = d\})}{card(\{R : x| = R\})}.$$

In formulas of aggregation factors, *max* denotes the maximum value, *avg* denotes the arithmetic average value, and *class*(R) denotes the class indicated by the rule R. To calculate the quality factor of a rule, we can use either the total support factor or the class support factor.
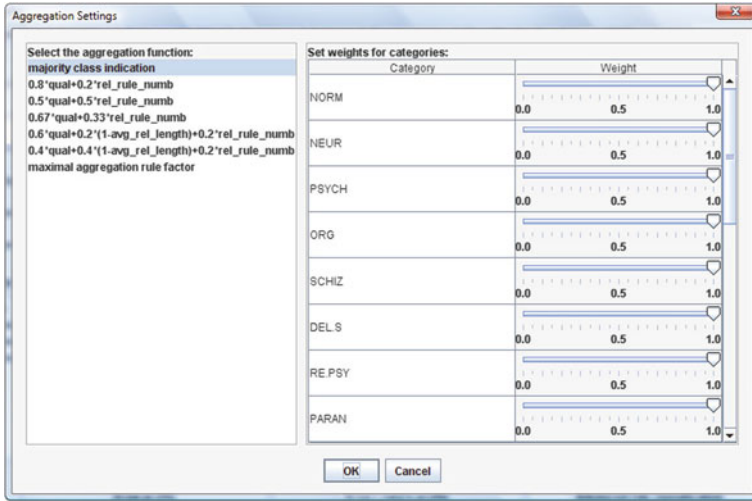
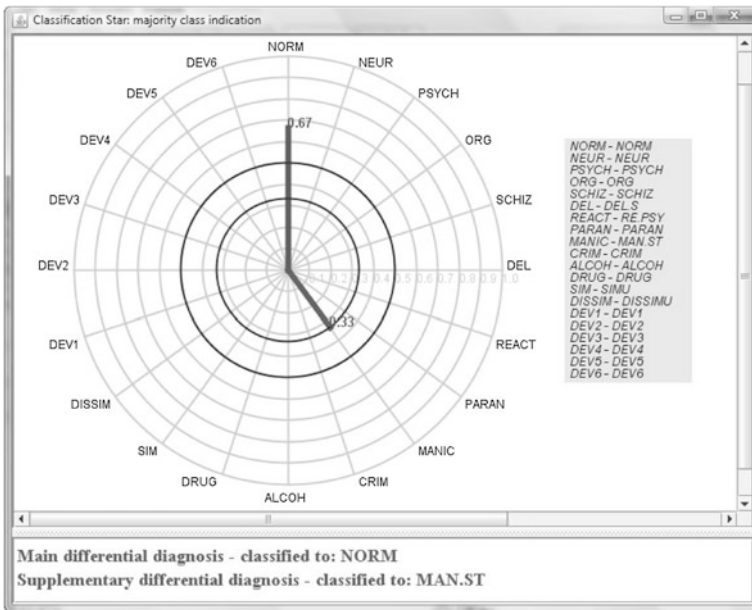**Fig. 4** Selection of aggregation factors in the Copernicus system

Fig. 5 An exemplary classification star in the Copernicus system

If a given aggregation factor $aggr(d)$ is calculated for each class d to which cases can be classified, then weighted maximum value is determined:

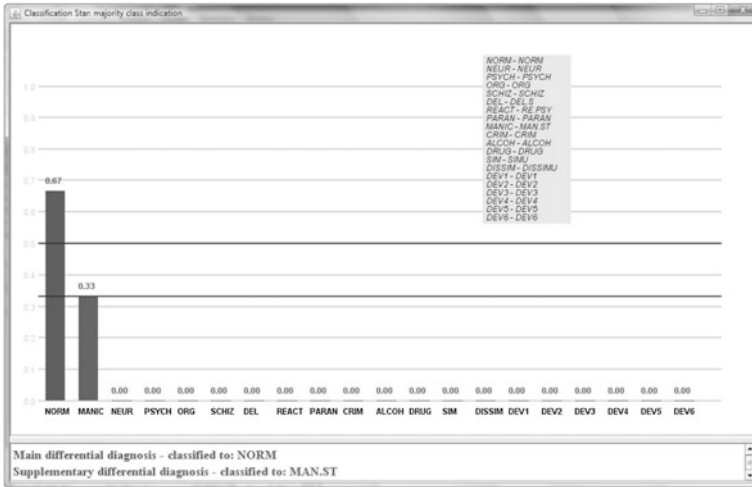$$\max(\{w_1 aggr(d_1), w_2 aggr(d_2), \ldots, w_m aggr(d_m)\}),$$

Fig. 6 An exemplary classification column chart in the Copernicus system

Table 3 A quality of classification of cases for described aggregation factors

Aggregation factor	Quality of classification (%)
$aggr_1$	87.9
$aggr_2$	77.4
$aggr_3$	84.8
$aggr_4$	82.9
$aggr_5$	78.9
$aggr_6$	82.6

where m is a number of all possible classes and weights $(w_1, w_2, \ldots, w_m)$ can be set by the user between 0 and 1, for each class separately (see Fig. 4).

The main differential diagnosis for a given case x is set as d_m if $w_m aggr(d_m)$ is the maximum value and $w_m aggr(d_m) > 0.67$. The supplementary differential diagnosis for a given case x is set as d_s if $w_s aggr(d_s)$ is the next maximum value and $w_s aggr(d_s) > 0.33$.

Classification results for each case are visualized in the form of the so-called classification star (see Fig. 5) or in the form of the so-called classification column chart (see Fig. 6).

Aggregation factors have been validated by experiments carried out on a data set with over 1,000 MMPI profiles of women. The quality of classification of cases for described aggregation factors is shown in Table 3. This quality is calculated as a ratio of a number of cases for which a class assigned by a diagnostician is the same as a class indicated by the classification system to a number of all cases.

4 Conclusions

In this chapter, we have described the Copernicus system—for computer-aided diagnosis of mental disorders based on personality inventories. The main attention has been focused on rule-based classification. This part of the tool has been presented more precisely. Our main goal of research is to deliver to diagnosticians and clinicians an integrated tool supporting the comprehensive diagnosis of patients. The Copernicus system is flexible and it can also be diversified into supporting differential diagnosis of profiles of patients examined by means of other professional multidimensional personality inventories.

References

1. Bazan, J.G., Szczuka, M.S.: The rough set exploration system. In: Peters, J., Skowron, A. (eds.) Transactions on Rough Sets III, Lecture Notes in Computer Science, vol. 3400, pp. 37–56. Springer, Berlin Heidelberg (2005)
2. Butcher, J. (ed.): MMPI: Research Developments and Clinical Application. McGraw-Hill Book Company (1969)
3. Choynowski, M.: Multiphasic Personality Inventory (in Polish). Psychometry Laboratory Polish Academy of Sciences, Warsaw (1964)
4. Cios, K., Pedrycz, W., Swiniarski, R., Kurgan, L.: Data Mining A Knowledge Discovery Approach. Springer, New York (2007)
5. Dahlstrom, W., Welsh, G.: An MMPI A Guide to use in Clinical Practice. University of Minnesota Press, Minneapolis (1965)
6. Dahlstrom, W., Welsh, G., Dahlstrom, L.: An MMPI Handbook, vol. 1–2. University of Minnesota Press, Minneapolis (1986)
7. Duch, W., Kucharski, T., Gomuła, J., Adamczak, R.: Machine learning methods in analysis of psychometric data. Application to Multiphasic Personality Inventory MMPI-WISKAD (in polish). Toruń (1999)
8. Gomuła, J., Paja, W., Pancerz, K., Mroczek, T., Wrzesień, M.: Experiments with hybridization and optimization of the rules knowledge base for classification of MMPI profiles. In: Perner, P. (ed.) Advances on Data Mining: Applications and Theoretical Aspects, LNAI, vol. 6870, pp. 121–133. Springer, Berlin Heidelberg (2011)
9. Gomuła, J., Paja, W., Pancerz, K., Szkoła, J.: A preliminary attempt to rules generation for mental disorders. In: Proceedings of the International Conference on Human System Interaction (HSI 2010). Rzeszów, Poland (2010)
10. Gomuła, J., Paja, W., Pancerz, K., Szkoła, J.: Rule-based analysis of MMPI data using the Copernicus system. In: Hippe, Z., Kulikowski, J., Mroczek, T. (eds.) Human-Computer Systems Interaction. Backgrounds and Applications 2. Part II, Advances in Intelligent and Soft Computing, vol. 99, pp. 191–203. Springer, Berlin Heidelberg (2012)
11. Gomuła, J., Pancerz, K., Szkoła, J.: Computer-aided diagnosis of patients with mental disorders using the Copernicus system. In: Proceedings of the International Conference on Human System Interaction (HSI 2011). Yokohama, Japan (2011)
12. Gomuła, J., Pancerz, K., Szkoła, J., et al.: Classification of MMPI profiles of patients with mental disorders—Experiments with attribute reduction and extension. In: Yu, J. (ed.) Rough

Set and Knowledge Technology, Lecture Notes in Artificial Intelligence, vol. 6401, pp. 411–418. Springer, Berlin Heidelberg (2010)

13. Gomuła, J., Pancerz, K., Szkoła, J.: Rule-based classification of MMPI data of patients with mental isorders: experiments with basic and extended profiles. Int. J. Comput. Intell. Syst. **4**(5) (2011)

14. Greenes, R.: Clinical Decision Support: The Road Ahead. Elsevier, Amsterdam (2007)

15. Lachar, D.: The MMPI: Clinical Assessment and Automated Interpretations. Western Psychological Services, Fate Angeles (1974)

16. Płuzek, Z.: Value of the WISKAD-MMPI test for nosological differential diagnosis (in polish). The Catholic University of Lublin (1971)

17. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann (2005)