Classification of MMPI Profiles using Decision Trees

Daniel Jachyra[1], Krzysztof Pancerz[1], and Jerzy Gomuła[2,3]

[1] Institute of Biomedical Informatics
University of Information Technology and Management in Rzeszów, Poland
djachyra@wsiz.rzeszow.pl
kpancerz@wsiz.rzeszow.pl
[2] The Andropause Institute, Medan Foundation, Warsaw, Poland
jerzy.gomula@wp.pl
[3] Cardinal Stefan Wyszyński University in Warsaw, Poland

Abstract. Our research concerns psychometric data coming from the Minnesota Multiphasic Personality Inventory (MMPI) test. MMPI is one of the most frequently used personality tests in clinical mental health as well as psychopathology (mental and behavioral disorders). We are developing the Copernicus system, a tool for the analysis and classification of MMPI profiles of patients with mental disorders. In this system, different quantitative groups of methods useful for differential interprofile diagnosis are selected and implemented. In the paper, we test different classification algorithms based on decision trees. Especially, for each algorithm, we take into consideration its ability to classify new cases. Results are important for selection of suitable algorithms to create a base for the Copernicus system.

Key words: classification, decision trees, computer-aided diagnosis, MMPI profiles

1 Introduction

Our research concerns psychometric data coming from the Minnesota Multiphasic Personality Inventory (MMPI) test [22]. MMPI is one of the most frequently used personality tests in clinical mental health as well as psychopathology (mental and behavioral disorders). The test builds upon multidimensional and empirical presumptions. It was designed and published first in 1943 in a questionnaire form by a psychologist S.R. McKinley and neuropsychiatrist J.Ch. Hathaway from the University of Minnesota. Later the inventory was adapted in above fifty countries. In our research, we have used data coming from the MMPI-WISKAD test. The MMPI-WISKAD personality inventory is a Polish adaptation of the American inventory. The test originally was translated by M. Choynowski (as WIO) [4] and elaborated by Z. Płużek (as WISKAD) in 1950 [26].

Between 1998 and 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

CONCURRENCY, SPECIFICATION AND PROGRAMMING
M. Szczuka et al. (eds.): Proceedings of the international workshop CS&P 2011
September 28-30, Pułtusk, Poland, pp. 397-407

MMPI-WISKAD test [7].Our current research is a continuation of those investigations. For two years, we have developed the Copernicus system (see [14], [13]). This is a system supporting clinical psychologists in differential and clinical diagnosis based on the overall analysis of profiles of patients examined by means of personality inventories (see Section 3). Various quantitative groups of methods useful for differential interprofile diagnosis have been implemented (cf. [11], [15]) using both supervised and unsupervised approaches [5].

In our previous works, we have tested different rule generation methods for their abilities to classify MMPI profiles (see [11], [15], [13], [16], [12]). The knowledge base embodied in the Copernicus system consists of a number of rule sets generated by different data mining and machine learning tools, such as:

- The Rough Set Exploration System (RSES) - a software tool featuring a library of methods and a graphical user interface supporting a variety of rough set based computations [2].
- WEKA - a collection of machine learning algorithms for data mining tasks [19], [30].
- The GhostMiner System [1] - a tool providing, among others, several different types of data mining algorithms, especially FSM Neuro-Fuzzy System and SSV Decision Tree.
- The STATISTICA Data Miner [20] - a part of the STATISTICA Data Analysis and Data Mining Platform.
- RuleSEEKER - a belief network and rule induction system [18].

In this paper, we focus on decision rules generated through decision trees. In Section 6, we present results of experiments carried out using several important decision tree algorithms implemented in the WEKA system. Selected algorithms will be embodied in the Copernicus system. Moreover, a visualization of classification of MMPI profiles using decision trees is designed and implemented in Copernicus (see Section 5).

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: $a(x) = [a_1(x), a_2(x), ..., a_{13}(x)]$. 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. For the training data (which are used to learn or extract relationships between data), we have a tabular form (see example in Table 1) which is formally called a decision system (decision table) $S = (U, A, d)$ in Pawlak's form [24]. U is a set of cases (patients), A is a set of descriptive attributes corresponding to scales, and d is a decision attribute determining a nosological type (class, category).

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),

Table 1. An input data for Copernicus (fragment)

Attribute	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	a_{13}	*class*
Scale	L	F	K	$1.Hp$	$2.D$	$3.Hy$	$4.Ps$	$5.Mf$	$6.Pa$	$7.Pt$	$8.Sc$	$9.Ma$	$0.It$	
#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*
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

$6.Pa$ (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.

In our research, we have obtained input data which have classes (nosological types) assigned to patients by specialists. Our data are categorized in two ways:

- Nineteen nosological classes and the reference class (*norm*). 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*).
- Six more general nosological classes called macroclasses: not-validated (*nval*), dependence (*depend*), organic (*org*), neuroticism (*neurot*), psychoticism (*psychot*), sociopathy (*socio*), and the reference class (*norm*).

3 The Copernicus System

The Copernicus system supporting clinical psychologists in differential and clinical diagnosis based on the overall analysis of profiles of patients examined by means of personality inventories is a tool designed for the Java platform. The main features of the application are the following: *multiplatforming* (thanks to the Java technology, the application works on various software and hardware platforms), *user-friendly interface* (the interface is designed in order to make it possible to use the application in the medical environment), *the module of data visualization* (it allows presenting data in a clear and comprehensible way, for example, in a graphical way for a person who must make a reasonable diagnostic decision), *modularity* (the project of the application and its implementation takes into consideration modularity in order to make it possible to extend the

application in the future and enlarge its usage on diagnosis based on different personal inventories). We can distinguish three main parts of of the Copernicus system:

- *Knowledge base.* The knowledge base embodied in the Copernicus system consists of: *classification functions*, *classification rule sets*, *decision trees*, *cluster characteristics*, *nosological category patterns.*
- *Multiway classification engine.* One of the main tasks of building expert systems is to search for efficient methods of classification of new cases. Classification in Copernicus is made on the basis of several methodologies. We can roughly group them into the following categories: rule-based classifiers (including decision-tree-based), distance-based classifiers, statistics-based classifiers. For each methodology, the most popular classifiers have been selected and implemented.
- *Visualization engine.* In the Copernicus system, a special attention has been paid to the visualization of analysis of MMPI data for making a diagnosis decision easier. The Copernicus system enables the user to visualize: classification rules in the form of stripes put on patients' profiles, classification functions, clusters of patients' profiles, decision tress with tracking decision paths for examined patients, specialized diagrams (e.g. Diamond's diagram, Leary's diagram), dendrograms.

4 Decision Tree Algorithms

In this section, we give a short characterization of decision tree algorithms implemented in the WEKA system [19], [30].

4.1 J48

J48 is an open source Java implementation of the C4.5 algorithm in the WEKA. This algorithm is used to generate a decision tree developed by R. Quinlan [27]. J48 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by J48 can be used for classification, and for this reason, J48 is often referred to as a statistical classifier. The J48 algorithm constructs a decision tree using the concept of information entropy. At each node of the tree, J48 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. Its criterion is the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision. The J48 algorithm then recurs on the smaller sublists.

4.2 LMT

The LMT (Logistic Model Trees) algorithm is a combination of tree induction and a logistic regression model resulting in a single tree [23]. A logistic model tree

consists of a standard decision tree structure with logistic regression functions at the leaves selected using posterior class probabilities. Therefore, LMT has a tree structure made up of a set of inner nodes and a set of leaves in an instance space. Tree induction identifies subdivisions of the instance space by recursively splitting it in a divide-and-conquer fashion until further subdivisions are not beneficial. In a study conducted on 36 datasets, Landwehr et al. concluded that the LMT model outperformed simple logistic, multi logistic, C4.5 and CART methods. In general, LMT model produces smaller trees than the classification trees built by C4.5 or CART.

4.3 RandomTree

Random Model Trees are essentially the combination of two existing algorithms in machine learning: single model trees are combined with Random Forest ideas [25]. Model trees are decision trees where every single leaf holds a linear model which is optimized for the local subspace described by this leaf. This works well in practice, as piece-wise linear regression can approximate arbitrary functions as long as the single pieces are small enough. The success and efficiency of Random Model Trees critically depends on some specific engineering features. Determining the best split point for an attribute is expensive: the data must be sorted according to this attribute, and then a linear scan can determine the best split for minimizing the weighted sum squared error.

4.4 REPTree

REP (Reduced Error Pruning) appears to be a very simple, almost trivial algorithm for pruning. There are many different algorithms that go under the same name. No consensus exists whether REP is a bottom-up algorithm or an iterative method. Neither it is obvious whether the training set or pruning set is used to decide the labels of the leaves that result from pruning [8].

4.5 J48graft

Grafting is an inductive process that adds nodes to inferred decision trees. This process is demonstrated to frequently improve predictive accuracy. Superficial analysis might suggest that decision tree grafting is the direct reverse of pruning. To the contrary, it is argued that the two processes are complementary. This is because, like standard tree growing techniques, pruning uses only local information, whereas grafting uses non-local information. The use of both pruning and grafting in conjunction is demonstrated to provide the best general predictive accuracy over a representative selection of learning tasks [29].

4.6 BFTree

A method used in the BFT (Best-First Tree) algorithm [28] adds the best split node to the tree in each step. The best node is the node that maximally re-

duces impurity among all nodes available for splitting (i.e. not labeled as terminal nodes). Although this results in the same fully-grown tree as standard depth expansion, it enables us to investigate new tree pruning methods that use cross-validation to select the number of expansions. Both pre-pruning and post-pruning can be performed in the way enabling a fair comparison between them. Best-First decision trees are constructed in a divide-and-conquer fashion similar to standard depth-first decision trees.

4.7 FT

Functional Trees (FT) are classification trees that could have logistic regression functions at the inner nodes and/or leaves. The algorithm that uses FT can deal with binary and multi-class target variables, numeric and nominal attributes and missing values [9].

4.8 NBTree

NBTree algorithm induces a hybrid of decision-tree classifiers and Naive-Bayes classifiers: the decision-tree nodes contain univariate splits as regular decision-trees, but the leaves contain Naive-Bayesian classifiers [21]. The approach retains the interpretability of Naive-Bayes and decision trees, while resulting in classifiers that frequently out-perform both constituents, especially in the larger databases tested. The induction of Naive-Bayes classifiers is extremely fast, requiring only a single pass through the data if all attributes are discrete. Naive-Bayes classifiers are also very simple and easy to understand.

4.9 SimpleCart

The CART (Classification and Regression Trees) algorithm was proposed by Breiman et al. [3]. A CART tree is a binary decision tree that is constructed by splitting a node into two child nodes repeatedly. The general aim of classification and regression tree analysis is the following. A set of observations and associated variables are given. The algorithm finds a way of using variables to partition the observations into homogeneously distributed groups, then use groups to predict observation. CART uses binary trees to recursively split observations with yes/no questions about variables.

5 Decision Trees in Copernicus

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 classification results have been designed and implemented.

A user can select a given decision tree generated in the WEKA format, translate it into the internal format of rules in the Copernicus system, and visualize the set of rules in the tabular form (see an example in Figure 1). In the standard versions of popular data mining systems, there is a lack of proper visualization tools, readable, for example, by diagnosticians. Next, a selected rule can be graphically presented as a set of stripes placed in the profile space. Each condition part is represented as a vertical stripe on the line corresponding to the given scale (see an example in Figure 2). Such visualization enables the user to determine easily which rule matches a given profile.

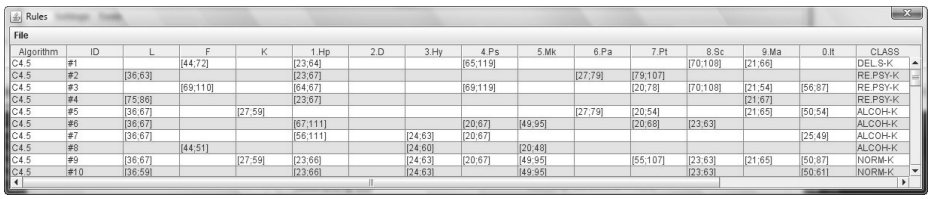

Fig. 1. A tabular form visualization of extracted rules from a decision tree (example)

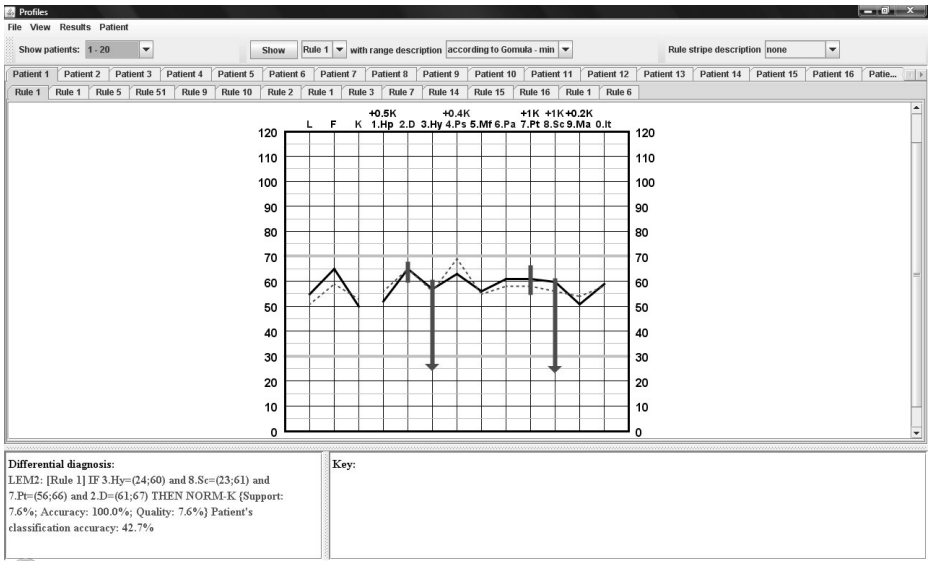

Fig. 2. Visualization of a rule in the profile space (example)

Decision trees coming form the WEKA system, and implemented in the Copernicus system enable diagnosticians to track a decision path (from the root to the leaf) for the selected patient's profile (see example in Figure 3).

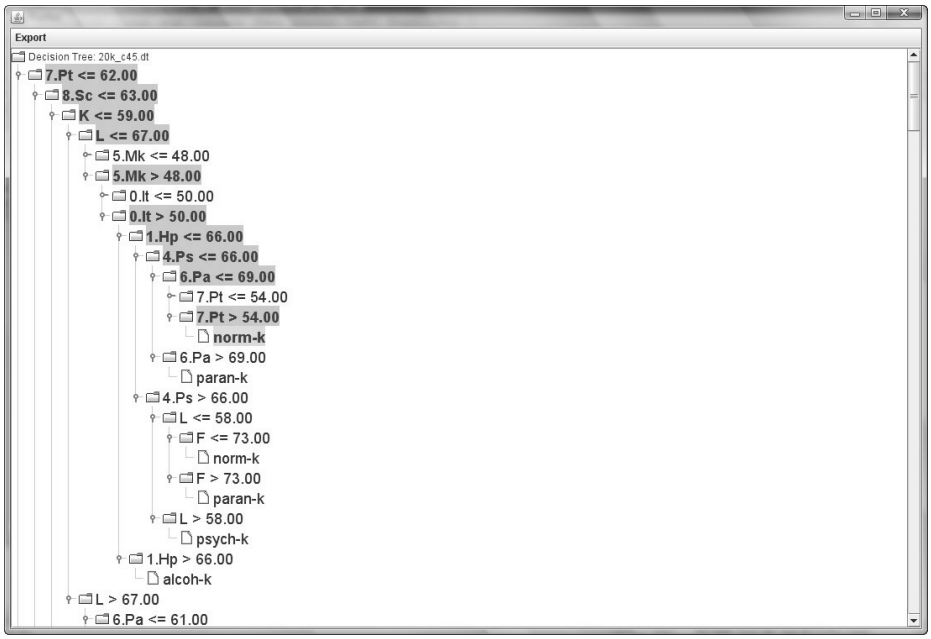

Fig. 3. Visualization of a decision tree with a decision path for a given patient's profile (example)

6 Experiments

In our experiments, decision tree algorithms mentioned shortly in Section 4 have been used. To determine the accuracy of classification of new cases a cross-validation method has been used. The cross-validation is frequently used as a method for evaluating classification models. It comprises of several training and testing runs. First, the data set is split into several, possibly equal in size, disjoint parts. Then, one of the parts is taken as a training set for the rule generation and the remainder (a sum of all other parts) becomes the test set for rule validation. In our experiments, the 10 cross-validation (CV-10) test was used.

Experiments were carried out on six data sets:

- 7w.arff - 7 macroclasses, women,
- 7m.arff - 7 macroclasses, men,
- 6wm.arff - 6 macroclasses, women+men,

– 20w.arff - 20 nosological classes, women,
– 20m.arff - 20 nosological classes, men,
– 20wm.arff - 20 nosological classes, women+men,

Table 2. Accuracy of classification

Data sets	J48	LMT	Random Tree	REPTree	J48graft	BFTree	FT	NBTree	Cart
7w.arff	93.22	94.38	89.80	90.20	93.28	92.12	90.84	91.94	92.17
7m.arff	90.04	91.72	89.48	86.96	89.87	88.36	87.80	90.10	88.64
6wm.arff	96.64	98.42	97.06	94.53	96.74	96.74	97.58	97.69	96.32
20w.arff	91.40	93.80	88.36	89.71	91.23	89.88	90.88	90.12	90.47
20m.arff	88.19	90.54	85.72	84.04	88.52	88.35	88.75	87.35	88.13
20wm.arff	89.32	92.50	89.80	87.70	89.32	88.51	91.03	90.94	89.11

In comparison to results obtained for rule-based systems (cf. [11], [15]), decision trees give better accuracy of classification. It is difficult to indicate distinctive algorithm in terms of classification accuracy. However, the LMT algorithm seems to outperform all others. One solution is to apply hybridization and optimization processes of the rule sets (see [10]). We could reduce a number of rules and simplify rules by removing some conditions from their predecessors. In the majority of cases, this process raises the classification accuracy. Rule generation algorithms very often give us superfluous rules, which can be removed.

7 Conclusions

In this paper, we have described a part of the Copernicus system - a tool for computer-aided diagnosis of mental disorders based on personality inventories. The main attention has been focused on decision-tree-based classification. Our main goal is to deliver to diagnosticians and clinicians an integrated tool supporting the comprehensive diagnosis of patients with mental disorders.

Acknowledgments

This paper has been partially supported by the grant from the University of Information Technology and Management in Rzeszów, Poland.

References

1. The GhostMiner System, `http://www.fqs.pl/business_intelligence/products/ghostminer`
2. Bazan, J.G., Szczuka, M.S.: The Rough Set Exploration System. In: Peters, J., Skowron, A. (eds.) Transactions on Rough Sets III, LNAI, vol. 3400, pp. 37–56. Springer-Verlag, Berlin Heidelberg (2005)

3. Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and Regression Trees. Chapman & Hall, Boca Raton (1993)
4. Choynowski, M.: Multiphasic Personality Inventory (in polish). Psychometry Laboratory, Polish Academy of Sciences, Warsaw (1964)
5. Cios, K., Pedrycz, W., Swiniarski, R., Kurgan, L.: Data mining. A knowledge discovery approach. Springer, New York (2007)
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. Elomaa, T., Kääriäinen, M.: An analysis of reduced error pruning. Journal of Artificial Intelligence Research 15, 163–187 (2001)
9. Gama, J.: Functional trees for classification. In: Cercone, N., Lin, T.Y., Wu, X. (eds.) Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM 2001), pp. 147–154. IEEE Computer Society (2001)
10. 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-Verlag, Berlin Heidelberg (2011)
11. Gomuła, J., Paja, W., Pancerz, K., Szkoła: A preliminary attempt to rules generation for mental disorders. In: Proceedings of the International Conference on Human System Interaction (HSI 2010). Rzeszów, Poland (2010)
12. 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 2. Advances in Intelligent and Soft Computing, Springer-Verlag, Berlin Heidelberg (2011), (to appear)
13. Gomuła, J., Pancerz, K., Szkoła: 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)
14. Gomuła, J., Pancerz, K., Szkoła, J.: Analysis of MMPI profiles of patients with mental disorders - the first unveil af a new computer tool. In: Grzech, A., Świątek, P., Brzostowski, K. (eds.) Applications of Systems Science, pp. 297–306. Academic Publishing House EXIT, Warsaw, Poland (2010)
15. Gomuła, J., Pancerz, K., Szkoła, J.: Classification of MMPI profiles of patients with mental disorders - experiments with attribute reduction and extension. In: Yu, J., et al. (eds.) Rough Set and Knowledge Technology, LNAI, vol. 6401, pp. 411–418. Springer-Verlag, Berlin Heidelberg (2010)
16. Gomuła, J., Pancerz, K., Szkoła, J.: Rule-based classification of MMPI data of patients with mental disorders: Experiments with basic and extended profiles. International Journal of Computational Intelligence Systems (2011), (to appear)
17. Greenes, R.: Clinical Decision Support: The Road Ahead. Elsevier (2007)
18. Grzymala-Busse, J., Hippe, Z., Mroczek, T.: Deriving belief networks and belief rules from data: A progress report. In: Peters, J., Skowron, A. (eds.) Transactions on Rough Sets VII, LNCS, vol. 4400, pp. 53–69. Springer-Verlag, Berlin Heidelberg (2007)
19. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: An update. SIGKDD Explorations 11 (2009)
20. Hill, T., Lewicki, P.: STATISTICS Methods and Applications. StatSoft, Tulsa, OK, USA (2007)

21. Kohavi, R.: Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In: Simoudis, E., Han, J., Fayyad, U.M. (eds.) Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD 1996). pp. 202–207. AAAI Press (1996)
22. Lachar, D.: The MMPI: Clinical assessment and automated interpretations. Western Psychological Services, Fate Angeles (1974)
23. Landwehr, N., Hall, M., Frank, E.: Logistic model trees. In: Lavrac, N., Gamberger, D., Todorovski, L., Blockeel, H. (eds.) Machine Learning: ECML 2003, LNAI, vol. 2837, pp. 241–252. Springer-Verlag, Berlin Heidelberg (2003)
24. Pawlak, Z.: Rough Sets. Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht (1991)
25. Pfahringer, B.: Random model trees: an effective and scalable regression method. Computer Science Working Papers 03/2010, Hamilton, New Zealand: University of Waikato, Department of Computer Science (2010)
26. Płużek, Z.: Value of the WISKAD-MMPI test for nosological differential diagnosis (in polish). The Catholic University of Lublin (1971)
27. Quinlan, J.: C4.5: Programs for Machine Learning. Morgan Kaufmann (1992)
28. Shi, H.: Best-first decision tree learning (2007)
29. Webb, G.I.: Decision tree grafting from the all-tests-but-one partition. In: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI 1999). pp. 702–707. Morgan Kaufmann (1999)
30. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann (2005)