Computational intelligence methods for information understanding and information management

Włodzisław Duch[1,2], Norbert Jankowski[1] and Krzysztof Grąbczewski[1]
[1]Department of Informatics, Nicolaus Copernicus University, Torun, Poland, and
[2]Department of Computer Science, School of Computer Engineering, Nanyang Technological University, Singapore

Abstract: Information management relies on knowledge acquisition methods for extraction of knowledge from data. Statistical methods traditionally used for data analysis are satisfied with predictions, while understanding of data and extraction of knowledge from data are challenging tasks that have been pursued using computational intelligence (CI) methods. Recent advances in applications of CI methods to data understanding are presented, implementation of methods in the GhostMiner data mining package [1] developed in our laboratory described, new directions outlined and challenging open problems posed. To illustrate the advantages of different techniques, a single dataset is exposed to the many-sided analysis.

Keywords: data understanding, knowledge extraction, decision support, data mining, computational intelligence, machine learning, neural networks, feature extraction, decision trees.

I. Introduction

Information management requires knowledge; abundant data about financial transactions, credit card use or state of sensors is available, taking gigabytes or even terabytes of memory. There is rough understanding of factors that influence formation of weather fronts, changes of the stock market shares or patterns of credit card use that indicate fraud. Discovery of these important factors, or understanding of the data, is most important, giving justification and powerful summarization of the knowledge contained in large databases. Although the main goal of statistics is to derive knowledge from empirical data, in most cases traditional statistical approaches are based on predictive models that do not give simple and satisfactory explanations of data structures. The same is true for most new data mining approaches based on computational intelligence. Fashionable predictive models based on Support Vector Machines [2], neural networks, neurofuzzy systems or evolutionary optimization [3][4] rarely generate knowledge that is understandable, that humans may learn and use in systematic reasoning. For example, prediction of the global climate change is very important, but understanding of the factors that facilitate the changes is of even greater importance. These changes could be summarized by a simple rule: IF fossil fuels are burned THEN climate warms up.

This paper presents a short review of methods aimed at data understanding developed in our laboratory, most of them contained in the data mining software package called GhostMiner [1]. In the next section several approaches to data understanding are presented, followed by a detailed exposition of these approaches. Summary of methods included in the latest GhostMiner package, some results and new developments conclude this paper.

II. Data understanding

Statistical methods, such as the Naive Bayesian, linear discrimination or their modern version based on kernel methods provide nonlinear hypersurfaces for data classification, while multilayered perceptron (MLP) neural networks combine many sigmoidal basis functions adjusting internal parameters to create, using training data, vector mappings from the input to the output space. All these methods allow for approximation and classification of data, enabling decision support in various applications but giving little understanding of the data [3][4]. As a result combining predictive models with *a priori* knowledge about the problem is usually difficult, many irrelevant attributes may contribute to the final solution. Large number of parameters in the data model, even in cases when data is abundant, may lead to overfitting and poor generalization. In novel situations, predictions of the black-box models may be quite unreasonable since there is no way to control and test the model in the areas of the feature space that are far from the training data. In safety-critical domains, such as medical, industrial, or financial applications, such risks may not be acceptable.

Much of the effort in CI approach to data understanding has been devoted to extraction of logical rules [5][6]. Reasoning with logical rules is more acceptable to human users than the recommendations given by black box systems, because such reasoning is comprehensible, provides explanations, and may be validated by human inspection. It also increases confidence in the system, and may help to discover important relationships and combination of features, if the expressive power of rules is sufficient for that. Searching for the simplest logical description of data with a large number of features is greatly simplified if good information selection techniques are used first, reducing dimensionality of the data. This reduction is always done with regard to the specific types of queries that the system for data analysis is designed for.

Symbolic description is not the only way to understand data. Human categorization is based on memorization of examples and creation of prototypes that are abstractions of these examples, rather than on logical rules defining natural objects in some feature spaces. “Intuitive understanding” is based on experience, i.e. on memorized examples of patterns combined with various similarity measures that allow for their comparison and evaluation. Prototypes searching may be combined with attribute selection methods, such combination simplify prototype understanding. Decision borders between different categories produced in this way may be quite complex and difficult to describe using linguistic statements. In symbolic Artificial Intelligence area this is explored in the case-based reasoning systems, but relatively little has been done in the computational intelligence community to create systems for data understanding based on prototype cases.

Visualization provides another way of understanding data, it forms a basis of the exploratory data analysis (EDA) that tries to uncover underlying data structure, detect outliers and anomalies, and find important variables [7]. Experts are able to understand the data simply by inspecting such visual representations. A special form of visualization is afforded by graphical methods that are aimed at the representation of the relationships between different elements of the problem description [8].

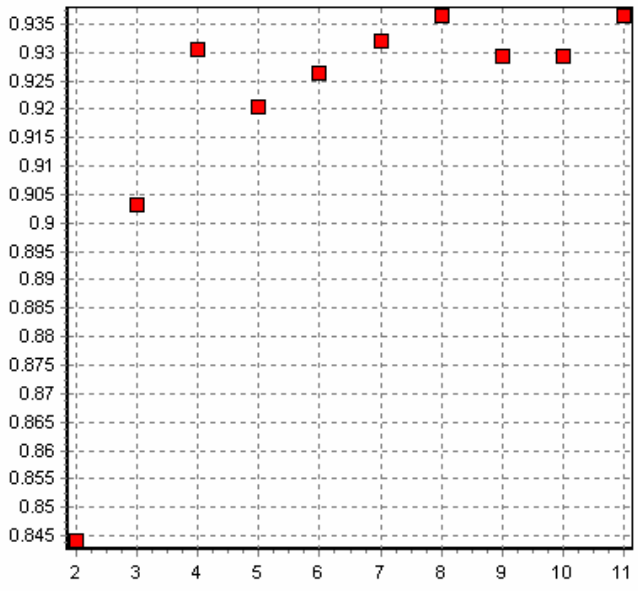
The best explanation of the data obviously depends on the type of the problem, the intention of user, as well as the type of questions and explanations that are commonly accepted in a given field. It is clear, however, that a good data mining system should provide many forms of data description. The GhostMiner software suite developed in our laboratory separates the process of data modeling, requiring some expertise in statistics and data modeling, from actual application of the data models to particular problems, performed by domain expert (a medical doctor, a marketing manager etc.). New data can be interpreted by trained models so that an expert in appropriate problem-field can analyze the case in detail.

III. Selection of information and dimensionality reduction.

Many databases we deal with have thousands or even hundreds of thousands of features. Such databases are quite common in text analysis, medical informatics or bioinformatics (see for example recent book [9] containing results from large-scale competition in information selection). Understanding the structure of such high dimensional data requires reduction of dimensionality, either selection of most valuable feature subsets, or aggregation of many features into new, more valuable ones. Decision support problems may frequently be formulated as classification problems, and features irrelevant to discrimination of particular classes may be filtered out. Many approaches to information filtering have recently been reviewed in [10]. A large library written in C++, called InfoSel++, implementing over 20 methods for feature ranking and selection, has been developed in collaboration with the Division of Computer Methods, Department of Electrotechnology, Silesian University of Technology (J. Biesiada, A. Kachel and T. Wieczorek). These methods are based on information theory (mutual information, information gain, symmetrical uncertainty coefficient, asymmetric dependency coefficients, Mantaras distance using transinformation matrix), distances between probability distributions (Kolmogorov-Smirnov distance, Kullback-Leibler distance), Markov blankets and statistical approaches (Pearson’s correlations coefficient (CC), Bayesian accuracy). These indices may be applied to feature ranking, ordering features according to their value. Information-theoretical and other indices depend very strongly on discretization procedures [11][12]. Therefore care has been taken to use unbiased probability and entropy estimators and to find appropriate discretization of continuous feature values.

Figure 1 presents an example of the influence of feature selection on the classification accuracy of 5NN model (*k* nearest neighbors, where *k* is equal to 5). The attribute selection performed here was simply taking the appropriate number of attributes which are most correlated with the class. The optimal points are placed between 6 and 11 attributes.

Figure 1. 5NN accuracy vs the number of features used.



This analysis (and most of the experiments presented in this paper) concerns the data taken from a Lancet article [15] (further referred to as the breast cancer data), where fine-needle aspirates of breast lumps were performed and age

plus 10 observations made by experienced pathologist were collected for each breast cancer case. The final determination whether the cancer was malignant or benign was confirmed by biopsy.

The SSV tree constructed on the breast cancer data is very simple, using single feature (the first PC) can classify with the accuracy of 93.7%. The appropriate factors of the first PC are -0.11, -0.17, -0.21, -0.31, 0.15, 0.01, -0.47, -0.51, -0.53, -0.17, 0.05.

Figure 2. Breast cancer dataset in the first two PCs.

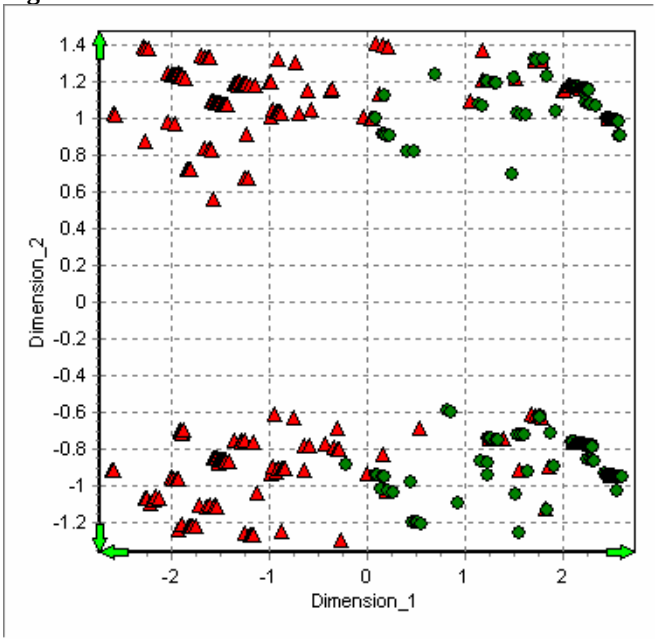


Figure 3. 5NN accuracy vs. number of attributes selected via SSV and passed to construct first PC.

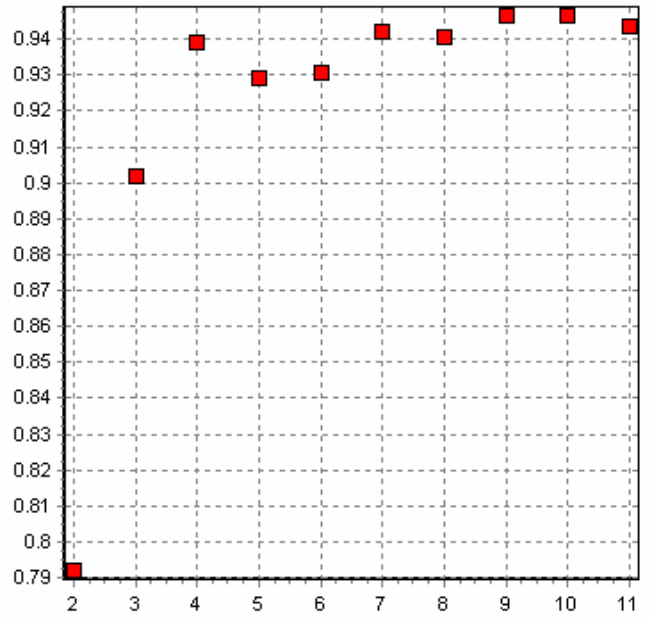
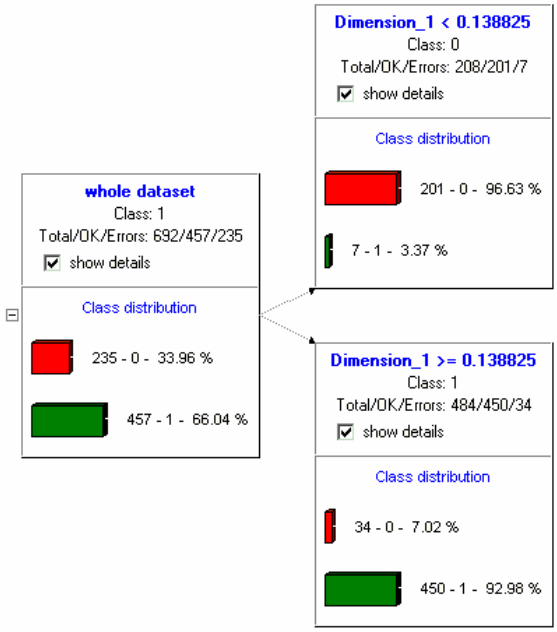


Figure 4. Simple SSV decision tree constructed from the first two PCs. Indeed it implies single logical rule.



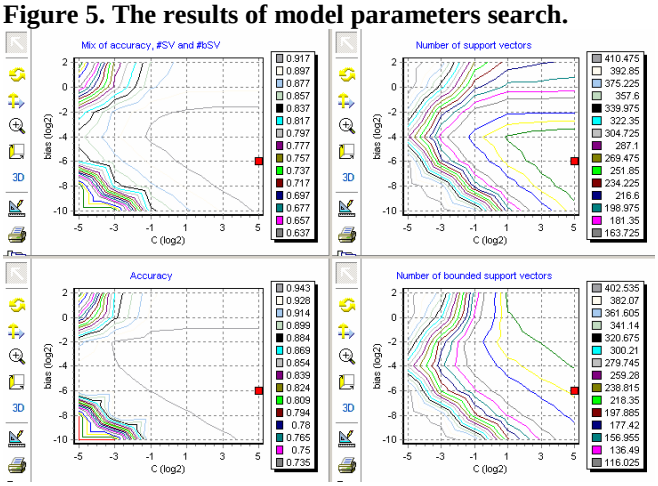
However before using PCA an attribute selection may be done as well. For example **Błąd! Nie można odnaleźć źródła odwołania.** presents how test accuracy depends on the number of attributes selected via SSV (this time used for attribute selection) and passed to PCA to extract first PC which after that were used to learn 5NN model. Results suggest that using four best attributes selected by SSV the PCA is constructed basing on four-dimensional input (in contrary to the previous example which used 11 attributes).

Many other methods of data transformation that reduce dimensionality are being experimented with. GhostMiner has no restrictions on the complexity of models that one may generate, thus one can use stacked models on features provided by initial models, for example each specializing in discrimination of a single class. An interesting and largely unexplored way of dimensionality reduction is based on multidimensional scaling (MDS) [14], an algorithm used primarily for data visualization. Using MDS for classification requires reduction of dimensionality of all data, both training and the query sets; this may be done because MDS is fully unsupervised procedure that does not use any information about class labels. However if the number of cases is big the minimization procedure become expensive. Unfortunately the meaning of features constructed in such a way is not easy to interpret.

Besides stacking of feature selection (or in general: data transformation) a selection of classification models can also be done with GhostMiner. Some classifiers are able to estimate not only *free* parameters but also may use inner configuration parameter estimation. For example the SVM model may estimate (sub-) optimal parameter C and the

spread of the Gauss kernel function (or parameter of other kernels, if used) – compare Figure 5. Moreover, optimal kernel may be estimated as well. In the case of kNN, the *k* (number of neighbors) may be estimated and/or automatic feature scaling can also be done.

Figure 5. The results of model parameters search.

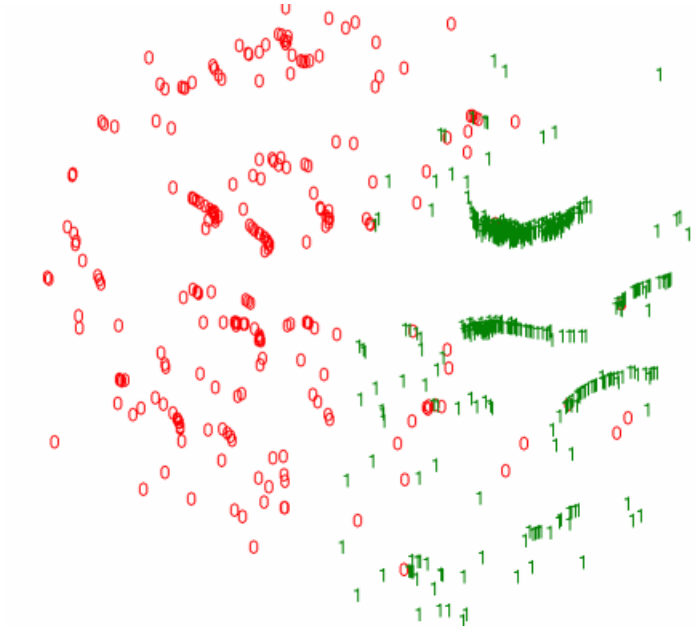


IV. Explanations based on visualization

Exploratory data analysis based on visualization is used as the first step. Scatterograms of pairs of interesting features show data distribution in two dimensions but this is frequently not sufficient. GhostMiner package contains two more sophisticated visualization methods. Principal components analysis (PCA) provides new features that may be used to present scatterograms for any pairs of induced features. MDS may use PCA results as the starting point in the procedure that tries to match the distances between the original high-dimensional data vectors and their representatives in the low-dimensional target space. Several types of MDS measures of topographical distortion are implemented in the GhostMiner package, with sophisticated way of performing required minimization [14].

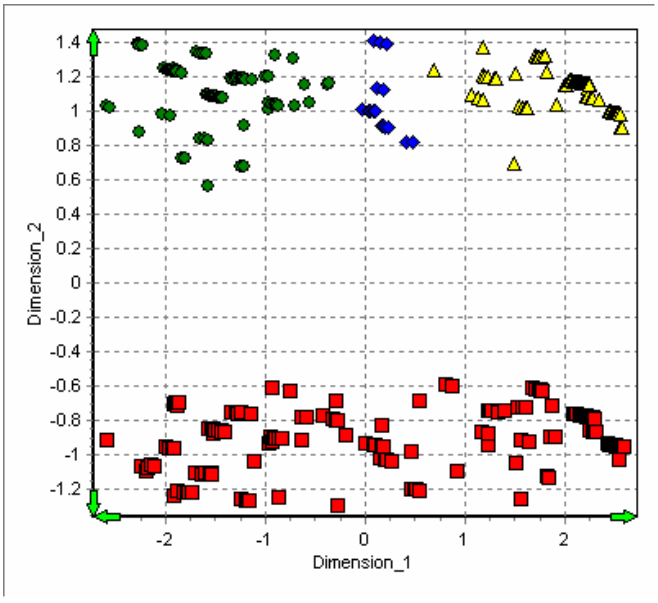
The MDS representation of 11-dimensional data (shown below) displays clear separation of the malignant (0) and benign (1) cases of cancer. The cluster structure shows much higher diversity of small clusters for the malignant case (left side), and several subtypes of benign cancer concentrated in a few clusters. Closer inspection shows that the prolonged structure of some clusters is due to the difference in the age of patients.

Figure 6. Two-dimensional MDS map of the breast cancer data.



Another visual data analysis may use clustering methods. It may be fruitful to compare results of clustering with data displayed with original class labels, especially using 2D scatterograms. Except dendrogram clustering a support vector (SV) clustering can be used in the GhostMiner system. The interesting feature of the SV clustering is irregular shapes – compare figure below.

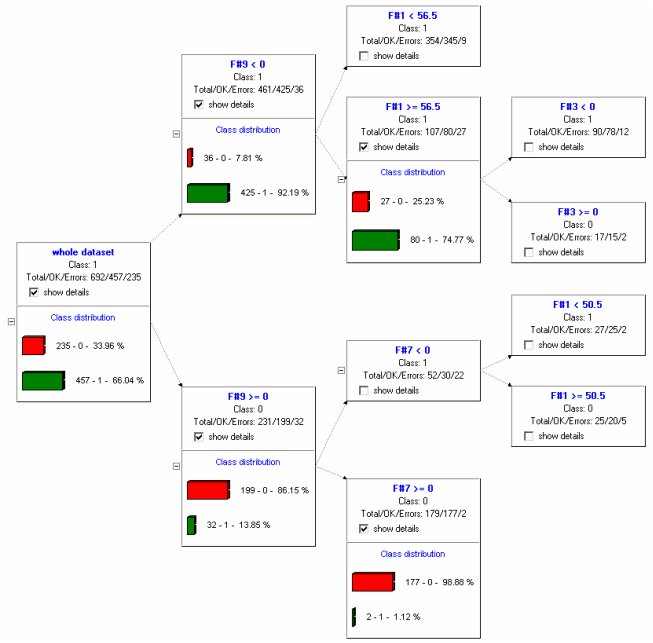
Figure 7. Support vector clustering of the breast cancer data.



V. Explanations based on logical rules

A very powerful way to summarize knowledge contained in databases is to extract logical rules from data. Our efforts to extract rules have been summarized in two longer papers [5][6]. In many cases univariate decision trees work quite well and because of their ease of use and computational efficiency they should always be used first. Our SSV tree, based on a separability criterion [13], provides several unique features. It may be used for feature discretization, feature ranking, and extraction of logical rules of different complexity. Crossvalidation techniques are used to define optimal parameters for pruning the tree branches, either by defining the degree of pruning or based on the leaf count. The classification trees can be constructed with best-first search or beam search. Applied to the breast cancer data, the method yields the decision tree presented in Figure 8.

Figure 8. SSV tree for the breast cancer data.



Some nodes have been expanded in figure below to show class proportions. The tree classifier achieves 95.4,% accuracy, with sensitivity of 90% and specificity of 98%. The presence of *necrotic epithelial cells* is identified as the most important factor. The decision function can be presented in the form of the following logical rules:

1. if F#9 > 0 and F#7 > 0 then class 0
2. if F#9 > 0 and F#7 < 0 and F#1 > 50.5 then class 0
3. if F#9 < 0 and F#1 > 56.5 and F#3 > 0 then class 0
4. else class 1.

With application of beam search, a simpler description can be obtained with slightly lower accuracy (94.8%), same sensitivity and specificity of 97%:

1. if F#9 > 0 and F#11 < 0 then class 0
2. if F#9 < 0 and F#1 > 56.5 and F#3 > 0 then class 0
3. else class 1.

Because rules differ significantly in their sensitivity and specificity it is worthwhile to create a "forest" of decision trees that have similar accuracy but rather different structure [16]. An analysis of such a forest generated for the breast cancer data reveals another interesting classification rules:

1. if F#8 > 0 and F#7 > 0 then class 0
2. else class 1.

Despite low accuracy (90.6%) the rule is valuable because of its simplicity and 100% specificity.

For many datasets a very simple and accurate knowledge has been generated using either the SSV trees or an MLP2LN algorithm [6] based on conversion of multilayer perceptron neural network into a logical network that performs function equivalent to a set of logical rules. The MLP2LN algorithm is not so easy to use as the SSV tree.

Combining logical rules with visualization gives more information, showing for example how typical the new case may be, how far from the border areas where decisions are not reliable. SVM models are not significantly more accurate in this case and do not provide much understanding of the data. The use of fuzzy rules that are generated using the Feature Space Mapping (FSM) algorithm built in the GhostMiner package does not improve results; neurofuzzy systems frequently generate large number of rules that are not comprehensible. In this case over 70 rules with Gaussian membership functions are generated, reaching similar accuracy as the crisp logical rules.

VI. Explanations based on prototypes

The GhostMiner software has several methods that allow it to find or create a useful prototype for a given category. One way to generate such prototype is to use decision tree with distance-based tests. Typical test in decision tree checks the split thresholds for individual features. We have also included tests based on distance to reference vectors. Surprisingly simple and accurate rules may be found in this way [16]. For the breast cancer data 4 prototypes give 96.4% accuracy, with sensitivity of 92.3% and specificity of 98.5%. Few other methods from out software package can find accurate solution even with just two prototypes, what simplify the total knowledge to two super-cases (with accuracy around 94%)! Such prototypes selection methods when used with feature selection in first stage reduce the

number of attributes from 11 to 6 without decrease of classification quality. More over, substituting the attribute selection by feature creation – taking first (only!) principal component also two prototypes are selected and final classifier keep similar precision. This means that finally model base on one feature and two prototypes. For a review on prototype selection methods see [19,20].

Prototype-based explanations are more general than neurofuzzy rules, although in many cases both approaches are equivalent [17]. For example, Gaussian membership functions in fuzzy rules correspond to the Euclidean distance functions. Optimization of positions of two prototypes are equivalent to linear discrimination, while adding more prototypes creates more complex, piecewise linear decision borders called Voronoi diagrams. Exactly the same space tesselation are generated by 1NN spread on the prototypes. On the other hand not all distance functions are additive and thus are equivalent to fuzzy rules. Probabilistic data-dependent distance functions may be applied to symbolic data and used to define natural membership functions in neurofuzzy systems. This allows to analyze data structures that cannot be described in feature spaces, but similarity between these data structures may be evaluated.

Relations between algorithms for training perceptrons and algorithms for prototype optimization have not yet been explored.

VII. Black box classification

To estimate the value of different explanations of given classification task, it is reasonable to see the relation between pure classification results of the comprehensive models and best classifiers available, even if they act as black boxes. To compare just classification accuracy a number of tests can be performed. One of the most reliable ways of accuracy estimation is crossvalidation. Multiple runs of the test facilitate better approximation of real generalization abilities of the models. Table 1 presents the classification results summary for the breast cancer data, evaluated in 10 repetitions of 10 fold crossvalidation of each model – the average accuracy (in percents) and standard deviation of the 10 crossvalidation results are presented.

Table 1 Summary of black box classification results.

Classifier	Average accuracy	Standard deviation
kNN	95,87	0,27
SVM	95,49	0,29
NRBF	94,91	0,29
SSV Tree	94,41	0,46
FSM	92,74	1,59
Naive Bayes	88,24	0,45

The k Nearest Neighbors model used here was equipped with automatic selection of most adequate value of the number of neighbors (k) – each training process performed a crossvalidation to select the best k in the range of 1-10. The SVM method summarized in the table used Gaussian kernels, but it must be pointed out that linear kernels also perform very well on this dataset leading to just 0.1% lower result of average accuracy. Although statistical significance of the differences between kNN or SVM and SSV Tree classifiers is quite high, the 1% decrease in accuracy can often be sacrificed for more comprehensive description of the model.

VIII. Conclusions

One may combine neural, fuzzy, similarity-based, rough, inductive, clustering, optimization, genetic and other evolutionary techniques in hundreds of ways to create data models, and to find and optimize sets of logical rules. Because there is an over-abundance of algorithms, only those computational intelligence techniques that proved to be directly useful to data understanding have been included in the GhostMiner software, but it already contains many methods that may be combined in a huge number of ways. Finding good models may require long and tedious experimentation. It seems that the most important step in development of data mining methods requires meta-learning techniques that will find all interesting methods for a given dataset automatically. Some steps in this direction have already been taken.

The *a priori* knowledge about a problem to be solved is frequently given in a symbolic, rule-based form. Extraction of knowledge from data, combining it with available symbolic knowledge, and refining the resulting knowledge-based expert systems is a great challenge for computational intelligence.

References

[1] GhostMiner, http://www.fqspl.com.pl/ghostminer/

[2] Scholkopf B., Smola A., *Learning with kernels*. MIT Press, Cambridge, MA, 2002

[3] Hastie T., Tibshirani R., and Friedman J., *The Elements of Statistical Learning*. Springer Verlag, 2001.

[4] Duda R.O., Hart P.E., and Stork D.G, *Pattern Classification*, New York: John Wiley & Sons, 2nd ed, 2001.

[5] Duch W, Setiono R, Zurada J.M, Computational intelligence methods for understanding of data. *Proc. of the IEEE* ,Vol. 92, No. 5, 771- 805, 2004.

[6] Duch W, Adamczak R, Grąbczewski K, A new methodology of extraction, optimization and application of crisp and fuzzy logical rules. *IEEE Trans. on Neural Networks* ,Vol. 12, No. 3, pp. 277-306, 2001.

[7] Jambu M, *Exploratory and Multivariate Data Analysis*. Boston, MA: Academic Press, 1991.

[8] Jordan M, and Sejnowski T.J, Eds. *Graphical Models. Foundations of Neural Computation*. Cambridge, MA: MIT Press, 2001.

[9] Guyon, I, Gunn S, Nikravesh M, and Zadeh L, *Feature Extraction, Foundations and Applications*. Springer Verlag, Heidelberg, 2005.
[10] Duch W, Filter methods, in [9].
[11] Duch W, Winiarski T, Biesiada J, Kachel, A, Feature Ranking, Selection and Discretization. International Conf. on Artificial Neural Networks (ICANN), Istanbul, Turkey, pp. 251-254, June 2003.
[12] Duch W, Wieczorek T, Biesiada J, Blachnik M, Comparison of feature ranking methods based on information entropy. Proc. of Int. Joint Conf. on Neural Networks (IJCNN), Budapest, Hungary, IEEE Press, pp. 1415-1420, July 2004.
[13] Grąbczewski K, and Duch W, The separability of split value criterion. 5th Conf. on Neural Networks and Soft Computing, Zakopane, Poland, June 2000, pp. 201-208.
[14] Naud A, *Neural and statistical methods for the visualization of multidimensional data*. PhD thesis, Dept. of Informatics, Nicholaus Copernicus University, Torun, Poland, April 2001, http://www.phys.uni.torun.pl/kmk/publications.html
[15] Walker AJ, Cross S.S, and Harrison R.F, Visualization of biomedical datasets by use of growing cell structure networks: a novel diagnostic classification technique. *Lancet* Vol. 354, 1518-1522, 1999.
[16] Grąbczewski K, and Duch W, Heterogenous forests of decision trees. *Springer Lecture Notes in Computer Science*, Vol. 2415, 504-509, 2002.
[17] Duch W, Blachnik M, Fuzzy rule-based systems derived from similarity to prototypes. *Lecture Notes in Computer Science*, Vol. 3316912-917, 2004.
[18] Duch W, Similarity based methods: a general framework for classification, approximation and association, *Control and Cybernetics*, Vol. 29, No. 4, 937-968, 2000.
[19] Jankowski N, and Grochowski M, Comparison of instances selection algorithms: I. Algorithms survey, In: Artificial Intelligence and Soft Computing, pages 598-603. Springer, June 2004.
[20] Grochowski M, and Jankowski N, Comparison of instances selection algorithms: II. Results and comments, In: Artificial Intelligence and Soft Computing, pages 580-585. Springer, June 2004
[21] Grąbczewski K and Jankowski N, Mining for complex models comprising feature selection and classification, In: Isabelle Guyon, Steve Gunn, Masoud Nikravesh, and Lofti Zadeh, editors, Feature Extraction, Foundations and Applications. Springer, 2005.

Włodzisław Duch heads the Department of Informatics, an independent, interdisciplinary unit at the Nicolaus Copernicus University in Torun, Poland, and is a visiting professor at Nanyang Technological University in Singapore. He received PhD in quantum chemistry (1980), DSc in computational physics (1986), worked in Japan, Canada, Germany, France and the USA; in 2003 after sabbatical in School of Computer Engineering, Nanyang Technological University, Singapore, he joined the faculty there as a Visiting Professor. He has written 3 books, co-authored and edited 4 books and wrote about 300 scientific papers.
For his full CV and many papers Google "Duch".

Norbert Jankowski received his MSc from the Computer Science Institute of the Wroclaw University in 1992, and his PhD from the Institute of Biocybernetics, Polish Academy of Sciences in 2000. He is the author of a book "Ontogenic neural network" and nearly 50 scientific papers.

Krzysztof Grąbczewski received his MSc from the Mathematics and Computer Science Institute of the Nicolaus Copernicus University in 1994, and his PhD from the Institute of Systems Research, Polish Academy of Sciences in 2003. He has authored or co-authored about 40 scientific papers.