Publishing House
AKAPIT

COMPUTER METHODS IN MATERIALS SCIENCE
Informatyka w Technologii Materiałów

Vol. 9, 2009, No. 2

ROUGH SET FORMALIZATION OF AN INFERENCE PROCESS IN A DOE SUPPORTING EXPERT SYSTEM

JACEK PIETRASZEK

Mechanical Engineering Department, Cracow University of Technology,
Al. Jana Pawła II 37, 31-864 Kraków, Poland
E-mail: pmpietra@mech.pk.edu.pl

Abstract

The purpose of the paper was an inference process's formalization – in terms of a rough sets theory – of an existing DOE (design of experiment) supporting expert system DAX. This was the first step in extending the system from determined decision tree and rules-based into fuzzy-logic based inference and considering uncertainty in other approaches than only probabilistic. The paper is primarily addressed to the developers and users of expert systems. The second addressees are statisticians preparing experimental designs.

Key words: design of experiment, experimental design, artificial intelligence, rough set theory, expert system, modelling

1. INTRODUCTION

In this paper a conceptual assumptions of the advising expert system DAX modifications are presented. The assumptions, related to a knowledge database and inference rules, are prepared as a start point to build a system extension for fuzzy description of uncertainty and related data analysis methods originated from artificial intelligence domain. The rough sets theory is proposed as a formalism of inference engine's extension. DAX system's structure and a necessary outline of the rough sets theory are described below.

2. USEFULNESS OF THE EXPERT SYSTEM DEVELOPMENT

In general experimental research are very expensive and their high costs should be limited by a proper applied theory of experiments (Design of Experiments – DOE). DOE methods guarantee valid experiments planning, proper selection of data analysis methods and mistake-free interpretation of results. The modern DOE is so complex and mathematically formalised that its implementation into investigation process is very difficult. DOE requires too narrow specialized experimenter's skills. An expert system advising on each of investigation's stages is the solution of the deadlock [8,21,25]. The expert system collects an appropriate information about a planned experiment from a user, transforms the information during a data process driven by a knowledge database and subsequently offers a prepared expertise: suggested experimental design, recommended data analysis methods and applicable software programs.

DOE approach requires very detailed description of investigated object, precisely selected approximation models and unambiguous optimisation criterion. Unfortunately, data available for an investigator are usually uncompleted, low numbered, sometimes inconsistent, metrologically and lexically imprecise. This characteristic is particularly accurate for passive experiments or data mining in archives data warehouses. As a results, the investigator is forced

ISSN 1641-8581

to arbitrary and subjective selection of experimental design, data validation procedure, approximation model, data analysis methods. It leads to investigation narrowing and at last to unintentional degradation of achieved results. The conclusions obtained from this investigation may be narrowed, biased and in utmost situation – completely mistaken.

The obtained results and conclusions are typically presented for an external body: technical or economical board. Their evaluation criteria are various, imprecise and less-defined; even mutually contradictory. The reports should be prepared with natural language, with data quantification and aggregation. The results presentations with imprecise but understandable linguistic variable and summarization are better than precise but incomprehensive long table of numbers.

There are strong differences between fuzzy and probabilistic approaches to uncertainty. This is the reason for collecting and formalisation of knowledge about application of artificial intelligence methods in DOE domain [22]. It is particularly related to:

a) the fuzzy description of uncertainty as an approach alternative to the probabilistic [11],
b) irregular experimental designs structurally optimised (intelligent designs, smart-designs) [19,20,23,27],
c) data pre-processing especially related to lexical vagueness [15],
d) forecasting model selection and identification including fuzzy least squares method [1,10,28], artificial neural networks [6,7,12,18,19,24], MARS approximations [9],
e) selection, identification and interpretation of fuzzy data statistics and fuzzy statistics of data [3,14],
f) results presenting and reporting with data quantification, aggregation and linguistic summarization [15].

The knowledge about these issues should be propagated in classic publishing forms but also as a specific advising system being compound of expert and e-learning systems equipped with an appropriate knowledge database.

Due to high costs and time consumption of new system building, an alternative approach was selected: extending of the existing DAX advising expert system [21,25].

3. DAX SYSTEM MODIFICATION AREAS

The advising expert system DAX supports design of experiments process. It was built in the years 2000-2002 in team leaded by Z. Polański [21,25]. The system knowledge database contains 720 typical experimental designs and inference rules for designs selection. During dialog session the system asks user about planned experiment and then a selection filter is constructed basing on obtained answers. An expertise is the results of the system's work: list of recommended experimental designs which fulfil formal requirements of planned experiment. The user may subsequently choose one of the recommended designs as a best fitting of his non-described requirements.

The present knowledge database contains experimental designs described and classified according to 16 features:

a) star point value ALFA – describing position of a star point in a design; allowable values are: non applicable, real value greater than 1,
b) recommended programs ZP – suggesting software programs for further data analysis; allowable values are: non suggestion, STATISTICA, Design-Expert, CADEX:ESDET, InDE-F, universal (Mathematica),
c) investigation aim CB – describing intentional aim of investigation; allowable values are: non applicable, model identification, screening without model identification, screening with model identification, empirical optimisation,
d) input factors characteristic XOB – classification of input factors; allowable values are: lack of information, numerical and independent, numerical and dependent, compound of both, categorical and numerical,
e) remarks COM – short verbal of coded remarks; allowable values are: lack of remarks, verbal remarks, specialized design for empirical optimisation, specialized intelligent (smart) design,
f) input factors quantification DWW – factors quantification characteristic; allowable values are: lack of information, consistent with the design, regular, Steinhaus'es, normative, orthogonal, Tschebyshev's, random, intelligent, arbitrary,
g) composite feature KP – information of design ability for staging; allowable values are: non-composite design, composite design,
h) replication outside centre R – information about existence of outside centre replications; allowable

values are: non existed in design, existed in design,

i) restricting functions WM – information about additional restrictions; allowable values are: lack of restrictions, constant sum condition (mixtures), individual restricting functions,

j) number of centre cases N0 – information about number of cases in the centre of design; allowable values are: non-negative integers,

k) number of levels NX – number of input factors values; allowable values are: integers greater than 1,

l) number of blocks B – maximum number of blocks which may be considered in the design; allowable values are: positive integers,

m) different cases number ND – number typically equal to number of design cases; it may be different in designs with explicitly specified replications; allowable values are: positive integers,

n) design cases number N – number of all explicitly specified design cases including explicitly specified replications; allowable values are: positive integers,

o) approximating model FOB – recommended function for data forecasting; allowable values are: lack of recommendation, indeterminable, linear, full quadratic, linear with two-way interactions, linear-quadratic, particular polynomial, cubic spline, reduced polynomial,

p) number of input factors I – number of factors considered in the specified experimental design; allowable values are: positive integers.

An experimental descriptor E is constructed during dialog session with a user. Each pair: question-answer transforms previous value of descriptor E and simultaneously, by rules retrieved from the knowledge database, drives the selection of the next question. The final descriptor obtained after the last question-answer pair processing is transformed into selection filter. The filter selects experimental designs into the final expertise basing on the design features values included in the knowledge database. The selection filter is created in SQL query language as very complex WHERE phrase of SELECT clause. The WHERE phrase is constructed by a conjunction of conditions related to particular design features. The conditions may be defined in different forms as selectors of particular values, ranged values (for numerical features) or sets (for categorical features).

The additional ranking of particular designs is not included in the above selection scheme. It means that all design included in the final expertise have the same weight.

The future modification of the DAX system should take into consideration two main areas. The first is a change of knowledge database structure. It is required to allow addition of new types of designs, approximators and analysis methods. On the other hand, it will allow implementation of imprecise characteristics. The second modification area is a change of dialog structures and inference rules. It is required to allow imprecise answers and simultaneous multi-branching decision processing. The final expertise will contain recommendations sorted by weights.

Due to deterministic rules implemented in the original DAX system inference engine, it seems it is better to transforms initial rule-based formalism into different form which will allow introduction of fuzzy uncertainty into internal descriptions [4,13] and into dialogs with a user. The Pawlak's rough sets [16] appears as a good solution for this stage formalism.

4. Rough sets formalism outline

In the early 80's Pawlak [16,26] proposed specific method of incomplete information description. A set which for a lack of information cannot be described unambiguously is limited by two sets: closure of superset named *upper approximation* and subset named *lower approximation*. Not going into details of strict formal description, the upper approximation may be described as a set of those elements which membership to considered set cannot be denied. The lower approximation may be described as a set of those elements which membership to considered set is certain. It is obviously that the difference of upper and lower approximations is a set of elements which membership is not certain but cannot be denied. The best phrase for this membership is 'may be'. This approach is conceptually very similar to Pedrycz's shadowed sets [17] proposed later although formalisms are completely different. The area of membership described with 'may be' is identified but without assigning the particular numerical value to this sentence like in Zadeh's fuzzy sets formalism [2,5,29].

The formal description of rough sets is included below. Let **U** is a set of objects *u* described with features (attributes) *q* belonging to a set of all features **Q**:

$$\mathbf{Q} = \{q_i |\ 1 \le i \le n\} \qquad (1)$$

where: n – number of features (attributes) describing elements from set **U**. Let $\mathbf{V}_\mathrm{q}$ is a space of feature q_i all values:

$$\mathbf{V}_q = \{\upsilon_q |\ \upsilon_q = q(x) \wedge x \in U\} \qquad (2)$$

and **V** is a space of all features all values:

$$\mathbf{V} = \bigcup_{i=1}^{n} \mathbf{V}_{q_i} \,. \qquad (3)$$

Let f is two-argument information function which results in a value of the particular feature q for the particular object x:

$$f : \mathbf{U} \times \mathbf{Q} \to \mathbf{V} \,. \qquad (4)$$

Due to notation convenience, the above information function may be treated as a information functions family for particular features:

$$f_x : \mathbf{Q} \to \mathbf{V} \,. \qquad (5)$$

The ordered four of defined above elements creates *information system SI*:

$$SI = (\mathbf{U}, \mathbf{Q}, \mathbf{V}, f) \,. \qquad (6)$$

If it is necessary to create a decision rule for the information system, the **Q** features set is decomposed into sum of two mutually separable subsets: conditional features **C** and decisive features **D**:

$$\begin{aligned} \mathbf{Q} &= \mathbf{C} \cup \mathbf{D} \\ \mathbf{C} \cap \mathbf{D} &= \varnothing \end{aligned} \qquad (7)$$

In this way, the ordered five is created:

$$DT = (\mathbf{U}, \mathbf{C}, \mathbf{D}, \mathbf{V}, f) \qquad (8)$$

and it is named *decision table DT*. If the conjunction of **C** features is satisfied by an object then appropriate values for features **D** are implied.

The modelling of incomplete information is realised by excluding of **P** subset from features **Q**, while the other features have values unknown or ignored. The relation of *P-nondistinguish* for two element x_1 and x_2 from **U** set is introduced:

$$\forall \mathbf{P} \subseteq \mathbf{Q} \ \ \forall x_1, x_2 \in \mathbf{U} \ \ .$$

$$\forall \mathbf{P} \subseteq \mathbf{Q} \ \ \forall x_1, x_2 \in \mathbf{U} \quad x_1 \tilde{P} x_2 \Leftrightarrow \forall q \in \mathbf{P} \ \ f_{x_1}(q) = f_{x_2}(q) \qquad (9)$$

This relation is reflex, symmetric and transitive. It means that it is equivalence relation which divides **U** set into separable class of abstraction:

$$\forall x \in \mathbf{U} \ \exists! [x]_{\tilde{P}} : [x]_{\tilde{P}} = \{y \in \mathbf{U} |\ x \tilde{P} y\} \,. \qquad (10)$$

With the above definition of abstraction classes, the $\tilde{P}$*–lower approximation* of **X** set from **U** space may be defined as a set of all abstraction classes for $\tilde{P}$ relation, included in **X** set:

$$\underline{\tilde{P}}\mathbf{X} = \{x \in \mathbf{U} : [x]_{\tilde{P}} \subseteq \mathbf{X}\} \qquad (11)$$

The above definition means that all abstraction classes creating lower approximation are certain included in **X** set. It means that the lower approximation is a set of element certainly belonged to **X** set.

$\tilde{P}$*–upper approximation* of **X** set from **U** space is a set of all abstraction classes for $\tilde{P}$ relation having non-empty union with **X** set:

$$\overline{\tilde{P}}\mathbf{X} = \{x \in \mathbf{U} : \ [x]_{\tilde{P}} \cap \mathbf{X} \ne \varnothing\} \qquad (12)$$

The above definition means that each of abstraction classes creating upper approximation *must* contain elements from **X** set and *may* contain elements non belonged to **X** set.

The difference of $\tilde{P}$–upper i $\tilde{\boldsymbol{P}}$–lower approximation of **X** set is named $\tilde{\boldsymbol{P}}$ *-boundary* area of **X** set and it is set of elements which membership in **X** set is not certain:

$$\mathbf{Bn}_{\tilde{P}}(\mathbf{X}) = \overline{\tilde{P}}\mathbf{X} \setminus \underline{\tilde{P}}\mathbf{X} \qquad (13)$$

The uncertainty modelled by rough sets is related to incompleteness of owned information, not to a immanent imprecise of its measure. Due to this aspect, a concept of rough set outlined here is particularly useful for multidimensionally described objects considered in expert systems. During a session with a user, the system creates a description of searched solution (expertise) gradually, question by question narrowing considered variety of possible solutions. Formally, it is iterative creation of two series: descending of upper approximations and ascending of lower approximations. The final expertise may be rough set or crisp (precise) set. In the first case, the expertise is a set of object which belongs to the searched set *certainly* or *possibly*. The second case is rare, of course. The case most often met is a set internally indefinable which means that the expertise contains object *possibly* belonged to the searched set (possible solutions), but not contains element *certainly* belonged to the searched set (certain solutions).

5. PROPOSAL OF DESCRIBING INFERENCE ENGINE WITH ROUGH SETS FORMALISM

During the session with a user, the expert system creates an experiment descriptor E. The creation process may be described as a limited series:

$$E_0, E_1, \ldots, E_i, \ldots, E_p \tag{23}$$

where: E_0 – initial descriptor containing no information about experiment, E_i – descriptor in i-th step of dialog.

After obtaining answer for the last question in the particular decision branch, the system creates the final experiment descriptor E_p. The E_p descriptor defines experimental design equivalence relation in the knowledge database of the system: the design p_1 is in relation E_p with the design p_2, if both designs are positively selected by a filter created from E_p descriptor. It divides designs set into equivalence classes. The designs from an equivalence class differs only in features insignificant from a E_p descriptor point of view. In this formalism, the system final expertise is as a lower approximation of E_p descriptor (treated as a set). The lower approximation is realised by a classes of E_p-*equivalent* designs.

Typically an expertise is definable what means that the expertise is non-empty but it is not simple long list of all design in the knowledge database. In rare cases an expertise may internally indefinable what means that there is no design in the knowledge database satisfying requirements of the described experiment.

The next step on the way to considering non-probabilistic uncertainty should be widening of formalism into cases of fuzzy rough sets. It should allow fluently transferring DAX system to the fuzzy inference and assigning weights to the particular elements of an expertise.

6. EXAMPLE OF CREATION PROCESS

The following iterative process is presented as an illustration of the proposed formalism. Due to limitations of two-dimensional pictures, the experiment descriptor is constructed by only two experiment properties: number of factors (independent variables) i and number of cases n. The example bases on a selection of two-level fractional factorial design.

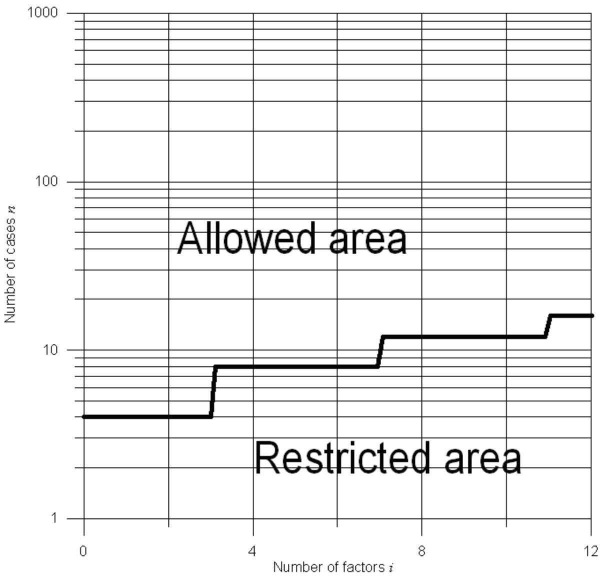

Fig. 1. *After imposing screening designs limitation.*

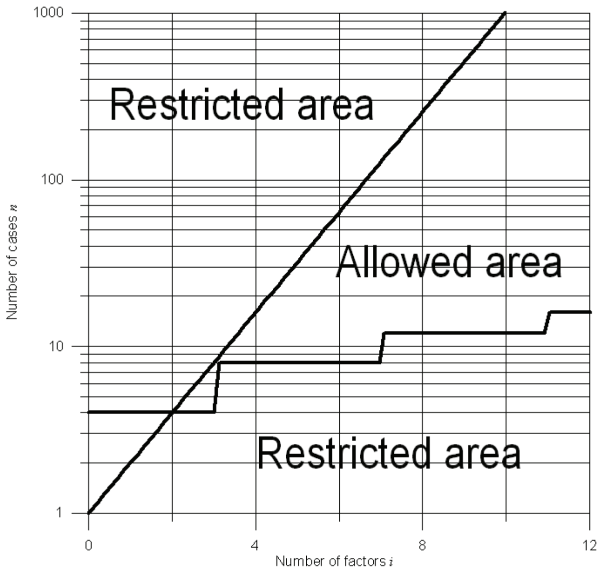

Fig. 2. *After imposing full factorial 2^i designs limitation.*

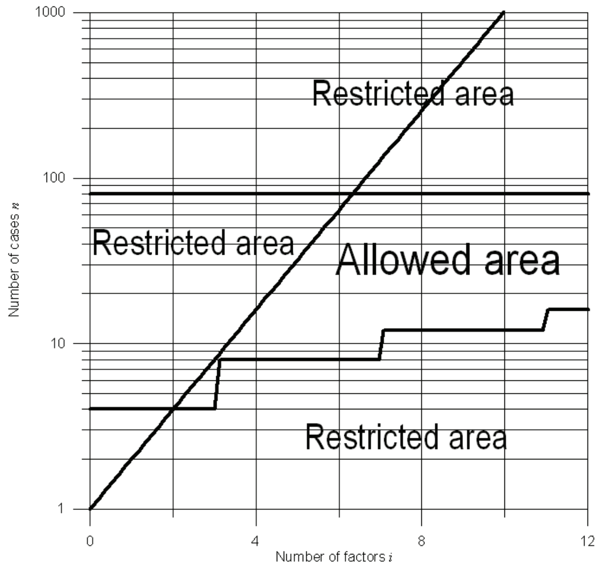

Fig. 3. *After imposing budget limitation ($n \leq 100$).*

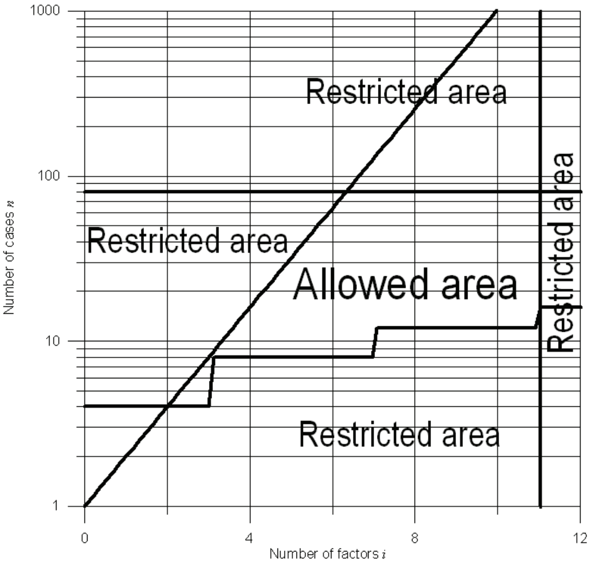

Fig. 4. After imposing factors limitation (i ≤ 11).

At first step, virtually the whole space is an expertise. It means that every experimental design is a member of the expertise. The first lower limitation is imposed by the minimum number of cases originated from population of screening designs (Plackett-Burman designs) and it divides the whole space in two subspaces presented in figure 1. A the second step, the upper limitation is imposed by a cardinality of full factorial 2^i design being presented in figure 2. At the third step, another upper limitation is imposed by a limited budget of investigation – the horizontal line in figure 3. At the fourth step, the next limitation is imposed by a maximum number of factors investigated – the vertical line in figure 4. At last, the area described as 'allowed' is an upper approximation of an expertise. Precisely: the expertise enumerates all experimental designs found in 'allowed' area. Also: nobody guarantees in general that an experimental design exists satisfying demands of the user.

Further, a membership function should be created to describe a level of belonginess. It will allow to introduce the fuzzy description to the expertise in the future.

7. CONCLUSIONS

The proposed modification of the advising expert system DAX is described in this paper. The aim of modification is to consider non-probabilistic approaches to uncertainty in a inference engine of the system. The usefulness of the work is argued: considering in the inference process approaches taken from artificial intelligence domain. The formalisation of inference process in the terms of rough sets theory is proposed. The simple example of descending rough sets sequence is included. The further planned work is outlined.

REFERENCES

1. Bárdossy, A., Hagaman, R., Duckstein, L., Bogardi, I., Fuzzy least squares regression: theory and application, Fuzzy Regression Analysis, eds, Kacprzyk, J., Fedrizzi, M., Omnitech Press, Warszawa, 1992, 181-193.
2. Bellman, R.E., Zadeh, L.A., Local and fuzzy logics, Modern uses of multiple valued logic, eds, Dunn J.M., Epstein D. , D.Reidel Verlag, Dordrecht, 1977, 103-165.
3. Buckley, J.J., Fuzzy statistics, Springer Verlag, New York, 2004.
4. Cubillo, S., Castiñeira, E., Bellido, S., On Fuzzy Inference by the Least Square Method for Non-Monotonic Functions, Soft Computing Foundations and Theoretical Aspects, eds, Atanassov, K.T., EXIT Press, Warszawa, 2004, 67-80.
5. Czogała, E., Pedrycz, W., Elementy i metody teorii zbiorów rozmytych, PWN, Warszawa, 1985 (in Polish).
6. Dobrzański, L.A., Maniara, R., Sokolowski, J., Kasprzak, W., Krupiński, M., Brytan, Z., Applications of the artificial intelligence methods for modeling of the ACAlSi7Cu alloy crystallization process, Journal of Materials Processing Technology, 192, 2007, 582-587.
7. Dobrzański, L.A., Kremzer, M., Trzaska, J., Włodarczyk-Fligier, A., Neural network application in simulations of composites Al-Al2O3 tribological properties, Archives of Material Science and Engineering, 30 (1), 2008, 37-40.
8. Dobrzański, L.A., Madejski, J., Prototype of an expert system for selection of coatings for metals, Journal of Materials Processing Technology, 175 (1-3), 2006, 163-172.
9. Friedman, J.H., Multivariate Adaptive Regression Splines, The Annals of Statistics, 19, 1991, 1-141.
10. Gładysz, B., Kuchta, D., Polynomial Least Squares Fuzzy Regression Models for Temperature, Artificial Intelligence and Soft Computing, eds, Cader, A., Rutkowski L., Tadeusiewicz R., Żurada J., EXIT Press, Warszawa, 2006, 118-124.
11. Grzegorzewski, P., Wspomaganie decyzji w warunkach niepewności. Metody statystyczne dla nieprecyzyjnych danych, EXIT Press, Warszawa, 2006 (in Polish).
12. Konieczny, J., Dobrzanski, L.A., Tomiczek, B., Trzaska, J., Application of the artificial neural networks for prediction of magnetic saturation of metallic amorphous alloys, Archives of Material Science and Engineering, 30 (2), 2008, 105-108.
13. Li, H.X., Yen, V.C., Fuzzy Sets and Fuzzy Decision-Making, CRC Press, Boca Raton, 1995.
14. Nahorski, Z., Horabik, J., Fuzzy Approximations in Determining Trading Rules for Highly Uncertain Emissions of Pollutants, Issues in Soft Computing. Theory and Applications, eds, Grzegorzewski, P., Krawczak, M., Zadrożny, S., EXIT Press, Warszawa, 2005, 195-209.
15. Niewiadomski, A., Interval-Valued Quality Measures for Linguistic Summaries, Issues in Soft Computing, Theory and Applications, eds, Grzegorzewski, P. Krawczak, M., Zadrożny, S., EXIT Press, Warszawa, 2005, 211-224.
16. Pawlak, Z., Rough sets, Intern. J. Comp. Inf. Sci., 11, 1982, 341-356.
17. Pedrycz, W., Shadowed sets: bridging fuzzy and rough sets, Rough Fuzzy Hybridization, A New Trend in Decision-

Making, eds, Pal, S.K., Skowron, A., Springer Verlag, Singapore, 1999, 179-199.

18. Pietraszek, J., Koncepcja meta-obiektu jako uogólnienie procesu aproksymacji neuronowej statycznych obiektów badań, Metrologia wspomagana komputerowo, ed., Przybysz, C., WAT, Warszawa, 3, 2001, 247-250 (in Polish).
19. Pietraszek, J., Response Surface Methodology at Irregular Grids Based on Voronoi Scheme with Neural Network Approximator, Neural Networks and Soft Computing, eds, Rutkowski, L., Kacprzyk, J., Springer-Physica Verlag, Heidelberg, 2003, 250-255.
20. Pietraszek, J., The criterion of homogeneity for space filling experimental design, The Improvement of the Quality, Reliability and Long Usage of Technical Systems and Technological Processes, eds, Boroszow, A.T., Bubulis, A., Silin, R.I., Royzman, W.P., Sokol, W.M., IFToMM Ukrainian Committee, Khmelnitsky, 2006, 24-28.
21. Pietraszek, J., Polański, Z., An expert system for computer aided selection of experimental designs, Proc. of 7th World Congress on Computational Mechanics (WCCM VII, 2006), Los Angeles, 2006 (CD ROM).
22. Pietraszek, J., Artificial Intelligence Methods in Design of Experiments, Proc. of 2nd Intn'l. Conf. on Modern Achievements of Science and Technology, Sept.25-Oct.2, 2008, Netanya, Israel, 2008, 118-124.
23. Polański, Z., Badania empiryczne – metodyka i wspomaganie komputerowe, Współczesna metrologia, zagadnienia wybrane, ed., Barzykowski J., WNT, Warszawa, 2004, 124-216 (in Polish).
24. Polański, Z., Pietraszek, J., Górecka, R., Aproksymacja neuronowa metodą sekwencyjnej modyfikacji danych, Proc. Sem. NeuroMet'99, Zastosowanie sztucznych sieci neuronowych w symulacji i sterowaniu procesami metalurgicznymi, ed,, Kusiak, J., Akapit Press, Kraków, 1999, 21-37 (in Polish).
25. Polański, Z., Pietraszek, J., Górecka-Polańska, R., System ekspertowy planowania i analizy eksperymentu, Metrologia wspomagana komputerowo, eds, Przybysz, Cz., WAT, Warszawa, 3, 2003, 11-16. (in Polish).
26. Rutkowski, L., Metody i techniki sztucznej inteligencji, PWN, Warszawa, 2005 (in Polish).
27. Skowronek, A., Optymalizacja procesu generowania elastycznych planów eksperymentu, Czasopismo Techniczne, 7, 1-I/2007, 63-74 (in Polish).
28. Tyrala, R., Linear Systems with Fuzzy Solution, Issues in Soft Computing, Theory and Applications, eds, Grzegorzewski, P., Krawczak, M., Zadrożny, S., EXIT Press, Warszawa, 2005, 277-288.
29. Zadeh, L.A., Fuzzy sets, Information and Control, 8, 1965, 338-353.

FORMALIZACJA PRZY POMOCY TEORII ZBIORÓW PRZYBLIŻONYCH PROCESU WNIOSKOWANIA DORADCZEGO SYSTEMU EKSPERTOWEGO PLANOWANIA DOŚWIADCZEŃ

Streszczenie

Celem niniejszej pracy było sformalizowanie – na podstawie teorii zbiorów przybliżonych – istniejącego systemu ekspertowego DAX wspierającego planowanie doświadczeń. Jest to pierwszy krok w stronę rozbudowy systemu od zdeterminowanego i ścisłego drzewa decyzyjnego i systemu regułowego w stronę wnioskowania rozmytego umożliwiającego uwzględnienie niepewności w ujęciu innym niż probabilistyczne. Praca jest skierowana przede wszystkim do projektantów i użytkowników systemów ekspertowych, a także do statystyków przygotowujących plany doświadczeń.

Submitted: October 15, 2008
Submitted in a revised form: December 1, 2008
Accepted: December 8, 2008