The problem of computer diagnostics of complex systems is one of the non-trivial tasks of modern information technology. Such systems are, for example, computer networks, automatic and/or automated control systems for complex technological objects, including related to complex problems of environmental protection, biology, etc. In pattern recognition, one of the major problems is forming subspaces of informative features, which only in the «ensemble» allow diagnosing the states of such systems with a high degree of reliability.

An effective approach to solving this problem based on the principles of inductive modeling of complex systems is proposed. The quality criterion for recognizing classes of patterns is formulated, which also makes it possible to evaluate the quality of the constructed ensemble of informative features.

As an example, the problem of constructing an ensemble of informative features represented by a binary code based on the data of an experiment to determine the hazard levels of some plant protection products is considered. Real primary data on plant protection products used in practice were applied to recognize the effect of certain characteristics on the so-called integrated «hazard indicator».

Comparative numerical estimates of the effectiveness of the proposed approach are given. In this case, there can be a fivefold gain in the amount of computations for a relatively small number of input features equal to 5 compared to the known algorithms of the class considered in the paper. It is shown that, from a practical point of view, the described algorithm has advantages over the known algorithms with brute-force search of feature subspaces in pattern recognition problems

Keywords: computer systems, computer diagnostics, pattern recognition, complex system, informative features

UDC 004.031.6: 621.3.07
DOI: 10.15587/1729-4061.2021.229756

IMPROVED ALGORITHM FOR MATCHED-PAIRS SELECTION OF INFORMATIVE FEATURES IN THE PROBLEMS OF RECOGNITION OF COMPLEX SYSTEM STATES

Volodymyr Osypenko
Doctor of Technical Sciences, Professor*
E-mail: vvo7@ukr.net
Borys Zlotenko
Doctor of Technical Sciences, Professor, Head of Department*
E-mail: zlotenco@ukr.net
Tetiana Kulik
Doctor of Technical Sciences, Associate Professor*
E-mail: t-81@ukr.net
Svitlana Demishonkova
PhD, Associate Professor*
E-mail: mashuk2007@ukr.net
Oleh Synyuk
Doctor of Technical Sciences, Professor**
E-mail: synoleg@ukr.net
Volodymyr Onofriichuk
PhD**
E-mail: volodymyronofriychuck@gmail.com
Svitlana Smutko
PhD, Associate Professor**
E-mail: svsmutko@gmail.com
*Department of Computer Engineering and Electromechanics
Kyiv National University of Technologies and Design
Nemirovycha-Danchenka str., 2, Kyiv, Ukraine, 01011
**Department of Machines and Apparatuses,
Electromechanical and Power Systems
Khmelnytskyi National University
Instytutska str., 11, Khmelnytskyi, Ukraine, 29016

Received date 25.02.2021
Accepted date 19.04.2021
Published date 30.04.2021

How to Cite: *Osypenko, V., Zlotenko, B., Kulik, T., Demishonkova, S., Synyuk, O., Onofriichuk, V., Smutko, S. (2021). Improved algorithm for selecting informative features with their paired accounting in the problems of recognizing the complex systems states. Eastern-European Journal of Enterprise Technologies, 2 (4 (110)), 48–54. doi: https://doi.org/10.15587/1729-4061.2021.229756*

1. Introduction

The problem of computer diagnostics of complex systems is one of the non-trivial tasks of modern information technology. Such systems include computer networks, complex technical architectures of information processing systems (more about the hardware part of such systems), automatic (automated) control systems for complex technological objects, complex electromechanical systems, etc. Systems related to complex problems of environmental protection, medicine, biology, agriculture are also complex systems that need constant diagnostics. This area includes a wide scientific and applied direction of pattern recognition, where one of the major problems is forming subspaces of informative features. Only in the «ensemble» these features allow performing the functions of diagnosing the states of such systems with a high

degree of reliability. However, real complex technical (technological) systems can be described by a large number of characteristics (parameters), which, in turn, may have a hierarchical structure. That is, one parameter may consist of several subparameters, which may have different effects on the state or behavior of the object under study.

In such cases, there is a problem of the need to work with multidimensional spaces of parameters (features – in terms of pattern recognition). Current (or predicted) or perhaps even critical states of modern complex systems can be described by multidimensional feature spaces $\{x_i \in X,\ i=1,2,...,n\}$. At the same time, the effectiveness of recognition (diagnosis) of a particular state may not necessarily depend on the entire available a priori set of features. Only a certain part of it can be decisive – the subspace of informative features for a given task – so-called «ensembles» of informative features $\{x_i^* \in X^* \subset X,\ i=1,2,...,n\}$. This means that precisely such features and precisely in such a «composition» make it possible to best recognize a situation that has developed in a complex technological object.

The solution to any recognition problem is directly related to the problem of finding a relevant feature system. Obviously, this is due to the huge variety of applied areas, which differ significantly in nature (physical, material, informational, biological, economic, etc.) and require the use of modern methods and tools of pattern recognition theory.

It is known that with the exhaustive search for all feature subspaces in an available input set, it is necessary to perform a fairly large amount of computations. For example, let the input set have 50 features and the maximum number of features in all possible options of feature subspaces would have a maximum of only 5 (n=5) parameters in the input data set. Then the number of options that need to be created and tested in order to select the optimal «ensemble» of informative features will reach 2,369,935 attempts. This, of course, will require numerous additional mathematical operations and, probably, the application of such an exhaustive search method in operational control problems of complex technological systems can be quite complicated and inconvenient.

Therefore, it is obvious that the problem of constructing (or selecting) feature subspaces in real applied pattern recognition problems remains very relevant today.

2. Literature review and problem statement

The effectiveness of solving a pattern recognition problem is usually evaluated by special quality (accuracy) criteria of recognition on a test data sample. A measure of feature informativeness can be a value that quantifies the ability of such a feature to recognize classes of patterns $k_r \in K$, where K is the number of clusters specified or formed in the process of recognition.

There are many approaches to assessing the informativeness of features, both in the practical and theoretical plane of pattern recognition theory. For example, [1] presents an approach to selecting a subset of features using a genetic algorithm. The feasibility of this approach for selecting a subset in the automated design of neural networks for pattern classification and knowledge discovery is demonstrated. A sequential scheme for selecting factors was applied. In [2], an attempt was made to apply self-organization methods to the problem of constructing a subset of features. However, the basic principles of computer self-organization of models in the construction of features are not applied. The paper [3] presents statistical criteria for assessing the informativeness of features of radiation sources of telecommunications networks and systems during their recognition. In [4], a rather wide set of statistical criteria for the features described by real numbers is presented. Evaluation here is also performed sequentially by the brute-force search for primary features, which, according to the authors, makes it possible to determine priorities of features and highlight the most informative ones. The paper [5] presents a semantic and [6] statistical approach to reducing spaces of input features to subspaces of their informative subsets. The works [7–9] are to some extent encyclopedic publications on pattern recognition in many areas in this powerful direction. Of course, the problems of constructing subsets of informative features for solving recognition problems are also covered.

It should be noted that these approaches can be quite effective with a small number of input features. For example, combinatorial or similar methods of search for all possible combinations of ensembles do well with the number of input features $n \le 20$. However, in some practical recognition problems, in particular and especially with binary descriptions, this number reaches hundreds and even thousands with limited capabilities of computer systems. The authors of [10] proposed an algorithm for selecting an ensemble of features, which applies the basic principles of the self-organization theory. This expanded the possibilities of selecting an ensemble of informative features compared to exhaustive combinatorial search, but since the advent of the algorithm described in [10], the complexity of problems, of course, has increased significantly.

Thus, further development of tools to reduce the amount and time of computation is important in the general field of pattern recognition theory.

3. The aim and objectives of the study

The aim of the study is to develop an improved algorithm for constructing an «ensemble» of informative binary features using the basic principles of inductive modeling of complex systems for pattern recognition problems.

To achieve the aim, the following objectives were set:

– to develop an improved algorithm for matched-pairs selection of informative features with step-by-step multi-row «selection» of intermediate results;

– to develop a criterion for assessing the quality of formed subspaces of informative features for specific problems;

– to conduct an experimental interpretation of the algorithm for forming «ensembles» of informative features to confirm its efficiency.

4. Research materials and methods

The research is based on a methodology that can be formulated as inductive modeling of complex systems (IMCS) based on input experimental data with interference. This methodology, in addition to many other applications, is aimed at solving pattern recognition problems in various fields and, particularly, in the field of innovative design of complex systems. The proposed algorithm uses the IMCS principles, in particular, the architecture of multi-row algorithms Group Method of Data Handling (GMDH) [11, 12] and this is its difference.

As is known [12], the IMCS methodology is based on three fundamental principles borrowed from different scientific fields, but organically created a holistic system of provisions. These principles can be formulated as follows:

1) the principle of heuristic self-organization, i. e. search for many candidate models and selection of the best ones by appropriately constructed so-called external model selection criteria («selection hypothesis»);

2) the principle of external compliment, i. e. the need to use «fresh information» in order to objectively verify models according to special criteria of regularity (accuracy);

3) the principle of freedom of choice and non-finality of decisions, or the principle of «freedom of choice of decisions for prospective open-end choice» [13, 14] – generation of not one but a set of intermediate results with the possibility to choose a subset of best options according to predetermined criteria.

Although the roots of such methods for solving recognition problems date back to the seventies and eighties of the twentieth century, for example [15], as of today they have been sufficiently developed in both theoretical and applied aspects. For example, the work [16] can be considered an encyclopedic collection of basic GMDH algorithms, including those that use schemes of multi-row inductive modeling algorithms. Most of these algorithms introduce the so-called structural-parametric identification of models of complex systems with automatic selection of subsets of informative parameters. The works [17, 18] also apply the principles of computer inductive modeling for clustering problems, including those that operate with large (several hundred, for example) dimensions of input feature spaces. However, the approaches presented in these works do not apply matching of features.

In general, practical applications of the IMCS methodology, in particular GMDH, have shown its effectiveness in various areas for problems with high levels of interference [12]. In this paper, this powerful direction of computer modeling in terms of using a multi-row architecture to build computational algorithms has also found direct application.

The computer experiment used the SELECT computer program developed by the authors, which implements a multi-row algorithm for matched-pairs selection of informative features. Some source materials for the experimental study of the proposed algorithm are taken from open sources – the statistical yearbook of the Food and Agriculture Organization of the United Nations (FAO) followed by processing (binarization).

5. Results of the study of the multi-row algorithm for matched-pairs selection of informative features

5. 1. Multi-row algorithm for matched-pairs selection of informative features

Classically, the decomposition of a general pattern recognition problem includes the following subtasks:

1) generating a set of primary features;

2) selection of a subset (subspace) of informative features;

3) construction of a decision rule or classifier;

4) assessment of recognition quality (usually on examination data samples).

In general, the problem of selecting informative features for further synthesis of decision (recognition) rules can be formulated as follows.

Let be a sample of input (a priori) data given as:

$$\left\{x_{ij} \in X, i = 1,2,...,n; j = 1,2,...,m\right\}, \quad (1)$$

where $\left\{x_{ij}\right\}$ is the array of values of input features of the object or process under study; i is the number of features in the set, j is the number of patterns (images, instances) in the given sample.

It is necessary to select the combination of features X^* from the original array of features X, which provides a minimum of a given evaluation criterion of the constructed set of features, which is conditionally written as:

$$Cr\left\{X^* \in X\right\} \to \min. \quad (2)$$

5. 1. 1. Criteria used

The proposed algorithm uses the so-called criterion «number of resolved disputes», which was formulated in [9], as the main one. This criterion allows distinguishing patterns at the information level.

Table 1 shows two classes of patterns R_1 and R_2. Note that a set of images that can be divided into (or which can objectively highlight) more than two classes can be reduced to a set with two classes. To do this, the first class R includes all images of some class, and the so-called «non-R class» $\overline{R}$ – all other images of the original set. Recognition of class R is carried out as if against the background of «non-R class» $\overline{R}$.

Table 1

Illustration of the optimization criterion «number of resolved disputes»

Class 1 (R_1)				Class 2 (R_2)			
Patterns	Features			Patterns	Features		
	x_1	x_2	x_3		x_1	x_2	x_3
ω_1	1	0	0	ω_4	1	1	1
ω_2	0	0	1	ω_5	0	1	0
ω_3	0	1	1	ω_6	1	0	1
$\omega^*_{R_1}$	0	0	1	$\omega^*_{R_2}$	0	1	0

Here ω_i (or ω_j) is the i-th (or j-th) image vector of the sample set; x_k is the k-th component (feature) of this vector; $\omega_i \in R$, $\omega_j \in \overline{R}$.

The «dispute resolution matrix» for the feature x_k is the following matrix:

$$X^k = \begin{pmatrix} a_{11} & a_{21} & ... & a_{n1} \\ a_{12} & a_{22} & ... & a_{n2} \\ ... & ... & ... & ... \\ a_{1m} & a_{2m} & ... & a_{nm} \end{pmatrix}. \quad (3)$$

where:

$$a_{ij} = \begin{cases} 1, \text{ if } x^k_{\omega_i} \neq x^k_{\omega_j}, \\ 0, \text{ if } x^k_{\omega_i} = x^k_{\omega_j}. \end{cases} \quad (4)$$

$$i = 1,2,...,n; \quad j = 1,2,...,m; \quad i \neq j; \quad k = 1,2,...,K.$$

In this example, for the features x_1, x_2, x_3, we have the following matrices:

$$X_1 \Rightarrow \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix}, X_2 \Rightarrow \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, X_3 \Rightarrow \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix}.$$

The multiplicity of dispute resolution is the value (denoted as $q_{\min}$) corresponding to the minimum term a_{ij} in (3).

The criterion «number of resolved disputes» in this case requires choosing the matrix (and hence the feature), where [8]:

$$q_{\min} \to \max. \quad (5)$$

That is, for some primary feature x_i, the dispute resolution matrix with greater $q_{\min}$ is better, which follows from natural prerequisites for solving recognition problems in the presence of interference.

As an additional criterion, the following one is expedient:

$$N(q_{\min}) \to \min, \quad (6)$$

which requires choosing the dispute resolution matrix, for which the number of elements corresponding to the multiplicity $q_{\min}$ is minimal. That is, it is actually a system of criteria for selecting the best options of feature subspaces:

$$\{q_{\min} \to \max \sim N(q_{\min}) \to \min\}. \quad (7)$$

In this example, from the constructed matrices for features x_1, x_2, x_3, it can be seen that the primary features do not allow distinguishing patterns of class R_1 from patterns of class R_2. This is explained by the fact that their matrices contain zero terms – unresolved disputes, i. e. there are zero elements in the corresponding «dispute resolution» matrices (3). The results can be improved by «overlapping» (a kind of element-by-element summation) matrices (3) by two, three, etc. In this case, the following options are possible:

1) $\{x_1, x_2\}$: $q_{\min} = 0$, $N(q_{\min}) = 3$;
2) $\{x_1, x_3\}$: $q_{\min} = 1$, $N(q_{\min}) = 0$;
3) $\{x_2, x_3\}$: $q_{\min} = 0$, $N(q_{\min}) = 2$;
4) $\{x_1, x_2, x_3\}$: $q_{\min} = 1$, $N(q_{\min}) = 0$.

The choice, obviously, will be in favor of the ensemble $\{x_1, x_3\}$, because for this solution, the values of the criteria $q_{\min} = 1$, $N(q_{\min}) = 0$:

$$x_1 x_3 \Rightarrow \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

In addition, the number of features in this ensemble to successfully recognize images of the two classes R_1 and R_2 is less ($n^* = 2$) than in the ensemble $\{x_1, x_2, x_3\}$ – ($n^* = 3$), for which the values of the criteria are the same. That is, the chosen ensemble allows solving the problem with a smaller number of features, where n^* corresponds to the optimal number of features in the informative ensemble $\{X^*\}$ for this problem.

5. 1. 2. Description of the multi-row matched-pairs feature selection algorithm

Unlike algorithms with brute-force search of dispute resolution matrices mentioned above, this algorithm performs their matching (search). Note that as a result of the algorithm, several matrices can be constructed in which the values of criterion (7) will be equal. In such a rare case, a matrix is chosen where the number of the next ascending value of $q_{\min}$ would be minimal. The algorithm consists of the following blocks.

*A*1 – rejecting obviously non-informative features. Non-informative features are those having the same value in all images, i. e. such that:

$$\sum_{j=1}^{m-1} \left(x_{j+1}^i - x_j^i\right) = 0, \quad i = 1,2,...,n, \quad (8)$$

where x_j^i is the i-th feature of the j-th image, m is the total number of images in the original set, n is the input number of features.

*A*2 – rejecting images with the same feature vectors as non-informative in advance.

*A*3 – constructing dispute resolution matrices for primary features $i = 1,2,...,\tilde{n}$, using rule (4), where $\tilde{n}$ takes into account possible exclusions of features in block *A*2.

*A*4 – constructing dispute resolution matrices for pairs of features $x_i x_l$ $(i = 1,2,...,\tilde{n}-1;\ j = i+1,2,...,\tilde{n})$.

*A*5 – selection of F best pairs of features according to criteria (5) and (6), taking into account the above remark (F – «freedom of choice of decisions» [13, 14]).

*A*6 – block of reassigning pairs of features according to the rule: $x_i x_l \Rightarrow y_k$.

*A*7 – checking the conditions for the values of criteria (6), (7). If $q_{\min}^s < q_{\min}^{s+1}$, then go to block *A*4, i. e. to the next selection row in terms of multi-row GMDH algorithms. In the case of equality $q_{\min}^s = q_{\min}^{s=1}$, but inequality $N(q_{\min}^s) > N(q_{\min}^{s+1})$, also go to *A*4; otherwise, go to the s-th selection row (block *A*4). Let d_α be the decision (conditional transition, where α corresponds to the block number) to be made. Then:

$$d_\alpha = \begin{cases} \alpha = 4, \text{if } q_{\min}^s < q_{\min}^{s+1}, \\ \alpha = 4, \text{ if } q_{\min}^s = q_{\min}^{s+1} \text{ and } N\left(q_{\min}^s\right) > N\left(q_{\min}^{s+1}\right), \\ \alpha = 8, \text{if } q_{\min}^s = q_{\min}^{s+1} \text{ and } N\left(q_{\min}^s\right) \le N\left(q_{\min}^{s+1}\right). \end{cases} \quad (9)$$

*A*8 – selection of the final ensemble of features. In the conditions of the last two blocks, the selection stop rule is set: on the last selection row, not F best pairs of features in terms of (5), (6) are chosen, but only one. Given block *A*6, it is possible to find the ensemble in terms of the input feature space. Thus, the constructed and selected ensemble X^* allows distinguishing images of class R^k against all other images of class $\bar{R}^k$.

Operations *A*3–*A*8 are repeated as many times as specified by the experts of classes in the input set of images.

5. 2. Quality assessment of the selected ensemble of informative features

Assessment of the informativeness of the obtained ensemble of informative features is made by the minimum of functionality, reflecting the accuracy of object recognition in the test sample. To decide whether the control sample ω_1^k belongs to a certain class R^k, in the training part of the sample, a decision rule is built in a perfect disjunctive normal form:

$$D_k = \sum_{\omega_i \in R^k} \varphi(X_k)_{\omega_i}, \quad k = 1,...,K, \quad (10)$$

where $\varphi(X_k)_{\omega_i}$ is the conjunction built for the selected ensemble X_k for the images $\omega_i \in R_k$.

Let $\Omega = \{\omega\}$ and $\Omega^* = \{\omega^*\}$ be training and control samples, respectively, and $R_k \subset \Omega$, $R_k^* \subset \Omega_k$. Then in the case of correct recognition of the k-th class, we have:

$$D_k = \begin{cases} 1, \text{ if } \omega^* \in R_k^*, \\ 0, \text{ if } \omega^* \notin R_k^*. \end{cases} \quad (11)$$

The criterion of recognition accuracy for the k-th class is written:

$$\varphi_k = \bar{D}_k + \bar{D}_k^*, \quad k = 1,...,K, \quad (12)$$

where $\bar{D}_k^*$ is the negation of the left part of (10), built on the set $R_k^* \subset \Omega_k$ for the class R_k^*.

Therefore, the functional (11) displays incorrect recognition for the selected ensemble.

The functional that displays correct recognition, and, therefore, characterizes the quality of the constructed ensemble of features, is as follows:

$$\Psi = \sum_{k=1}^{K} \varphi_k \to \min. \quad (13)$$

To illustrate the effectiveness of the quality criterion of the selected ensemble of informative features, which will display correct recognition, input binary data on physicochemical and toxic properties of substances can be used (Table 2).

Table 2

Input data on physicochemical and toxic properties of substances (binarized)

Physicochemical properties																				
No.	R_i	MW, X_{f1}					MP, X_{f2}					WS, X_{f3}					VL, X_{f4}				
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	R_1	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1
2		0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0
3		0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0
...	...	...																			
49	R_V	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0
50		0	0	0	1	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0

Toxic properties											
No.	R_i	LD_{50}, X_{f5}					AF, X_{f6}				
		21	22	23	24	25	26	27	28	29	30
1	R_1	0	1	0	0	0	0	0	0	1	0
2		0	0	0	1	0	1	0	0	0	0
3		0	1	0	0	0	0	1	0	0	0
...	...	...									
49	R_V	0	0	0	1	0	0	0	0	1	0
50		0	0	1	0	0	0	0	1	0	0

Let the selected ensemble be: $X_k = \{x_1, x_3\}$; ($q_{\min}$=0). Then:

$$D_1 = (x_1 \wedge \bar{x}_3) \vee (\bar{x}_1 \wedge x_3),$$

$$D_2 = (x_1 \wedge x_3) \vee (\bar{x}_1 \wedge \bar{x}_3). \quad (14)$$

Table 1 shows that the pattern ω_2 does not differ from ω_3, and ω_4 from ω_6, so the functions D_1 and D_2 contain not three but two conjunctions and

$$\Psi = \sum_{k=1}^{K=2} \varphi_k = (\bar{x}_1 \wedge x_3) \vee (\bar{x}_1 \wedge \bar{x}_3) = 0 + 0 = 0. \quad (15)$$

Thus, the system of rules (13) can recognize objects belonging to different classes by applying the already constructed ensemble of informative features $\{x_1, x_3\}$.

5. 3. Experimental application of the algorithm for selecting ensembles of informative features

An experiment to determine the hazard levels of some plant protection products on the basis of primary measurement data of the studied environment was considered. Measurement data come from special sensors through communication channels to the computer to recognize the impact of certain characteristics (features $x_i \in X$, $i=1,2,...,n$) on the integrated «hazard indicator» W, which can be described in the feature space $\{X\}$. Here the task is not to consider the purely technical side of the experiment, but to apply the above algorithm for selecting informative factors-features. Such indicators can be obtained within an automated environmental monitoring system to study those that most affect the value of W. This emphasizes that the described algorithm can be applied not only in the «technical» field, but also in other areas of research, such as in environmental studies or in health research with specified input data.

Input a priori information (pre-measured values of factors) is quite cumbersome, so Table 2 shows only a fragment of it with already binarized data. In Table 2, the following abbreviations for the properties of the substance are adopted: MW – molecular weight, MP – melting point, WS – water solubility, VL – volatility, LD_{50} – median lethal dose for white rats, AF – accumulation factor.

The numbers 1, 2,..., 29, 30 in Table 2 indicate the gradation numbers of the six properties with 5 levels for each. For example, for RV (water solubility), the number 11 (x_{11}) corresponds to the range of (0.01–0.02) g/l, 12 (x_{12}) – (0.03–0.04) g/l, etc., 15 (x_{15}) – (0.09–0.10) g/l. In the same way, the values of other features were obtained, but for each indicator from each property in its ranges and units.

Based on the results of the synthesis of the subsystem of features, a certain conclusion can be made for ecological and technical environmental monitoring from the standpoint of minimizing the negative impact of a particular product W (e. g., pesticide in agriculture) (Table 3).

Table 3

Results of selecting informative features for five classes of levels W

R_k	Physicochemical properties				Toxic properties	
	MW, X_{f1}	MP, X_{f2}	WS, X_{f3}	VL, X_{f4}	LD_{50}, X_{f5}	AF, X_{f6}
R_1	x_2	x_7	x_{11}	x_{16}	x_{25}	x_{26}
R_2	x_2	x_6	x_{11}	x_{20}	x_{24}	x_{26}
R_3	x_4	x_8	x_{11}	x_{16}	x_{24}	x_{27}
R_4	x_4	x_7	x_{12}	x_{16}	x_{24}	x_{28}
R_5	x_4	x_7	x_{13}	x_{16}	x_{22}	x_{28}

Table 3 shows that the most hazardous substances of class R_1 have high toxicity (x_{25}) and pronounced accumulation properties (x_{26}). For class R_1, on the contrary, low toxicity (x_{22}) and accumulation ability.

6. Discussion of the results of using the improved algorithm for matched-pairs selection of informative features

The described algorithm has advantages over known algorithms with brute-force search of feature subspaces in pattern recognition problems with large differences of input features represented by a binary code. This can be illustrated by the following example.

Suppose we have a problem of relatively low dimension, where the number of recognition classes k=2, the number of features n=16, the number of objects (images) l=16 and by one computational procedure we mean the computation of one element of the matrix (3).

With the exhaustive combinatorial search of all possible options, the number of operations is:

$$Q_1 = c^2 \sum_{i=1}^{n} C_n^i = 17 \cdot 10^8. \quad (16)$$

When using multi-row brute-force search [9], the number of operations is:

$$Q_2 = c^2\left[c + \sum_{i=1}^{n-1} c(c-i+1)\right] \approx 6\cdot 10^5. \quad (17)$$

When using the improved multi-row matched-pairs feature selection algorithm:

$$Q_3 = c^2\left(n + C_n^2(k) + C_F^2(k) + ... + C_F^2(k)\right) \approx 1.2\cdot 10^5,$$

$$k = 1,2,...,K; \quad K = \log_2(n-1). \quad (18)$$

Analysis of the written comparisons of the improved algorithm for matched-pairs selection of informative features with other algorithms aimed at solving the same problem, shows the following. With the exhaustive combinatorial search of all options of structures of feature subsets according to the given values of the input parameters k, n and l, it is necessary to search through the number of combinations represented by expression (16), which is $17\cdot 10^8$ operations. Thus, the application of the algorithm with multirow sequential matching of features under the same conditions and when a similar result is achieved allows performing this task for $6\cdot 10^5$ operations. That is, much faster, and the gain will be $S: S = Q_1/Q_2 \approx 5$ times.

The application of the improved multi-row matched-pairs feature selection algorithm to find the desired result according to expression (18) is estimated at about $1.2\cdot 10^5$ operations. This gives a gain of $S = Q_1/Q_3 \approx 130$ times in the amount of computations.

Such results are due to, firstly, matching of features and, secondly, the principle of multi-row selection of ensembles, inherent in the inductive approach to computer modeling of complex systems.

It should also be noted that these estimates for the second and third methods of selecting the resulting subset of features show the upper limits of the number of operations. In fact, these values may even be significantly lower. This is because the quality of ensembles of features is assessed on each selection row. The required set of informative features can be achieved earlier than would be required with the exhaustive combinatorial search of all options and their evaluation by the same criteria (6)–(8).

Although no loss of features from the primary information base during the multi-row procedure was found in test experiments, such a possibility exists and requires additional research in the future.

8. Conclusions

1. One of the approaches to solving the general problem of constructing subspaces of informative features presented by a binary code in the problems of recognition of complex system states is proposed. The algorithm uses the feature matching procedure, which indicates its effectiveness and makes it possible to significantly reduce the amount of computations.

2. Quality criteria for recognizing classes $q_{\min} \to \max$ and $N(q_{\min}) \to \min$, which also make it possible to assess the quality of the constructed ensemble of informative features in the system $\{q_{\min} \to \max - N(q_{\min}) \to \min\}$ are formulated. The model example shows the effect of such a criterion for recognizing two classes when using the selected ensemble of features.

3. An example of using the proposed algorithm to solve a specific practical problem of selecting an ensemble of informative features with an assessment of the effectiveness of such an ensemble in the examination sample is given. From a practical point of view, the described algorithm has advantages over known algorithms with brute-force search of feature subspaces in pattern recognition problems. This is shown for a relatively small (number of recognition classes $k=2$, number of features $n=16$, number of objects (images) $l=16$) problem. In this direction, we can conclude that the efficiency of such an algorithm will increase with increasing dimension of the feature space.

References

1. Yang, J., Honavar, V. (1998). Feature Subset Selection Using a Genetic Algorithm. Feature Extraction, Construction and Selection, 117–136. doi: https://doi.org/10.1007/978-1-4615-5725-8_8
2. Carpenter, G. A., Grossberg, S. (1987). ART 2: self-organization of stable category recognition codes for analog input patterns. Applied Optics, 26 (23), 4919. doi: https://doi.org/10.1364/ao.26.004919
3. Ilnitskiy, A., Burba, O. (2019). Statistical criteria for assessing the informativity of the sources of radio emission of telecommunication networks and systems in their recognition. Cybersecurity: Education, Science, Technique, 1 (5), 83–94. doi: https://doi.org/10.28925/2663-4023.2019.5.8394
4. Zayats, V. M., Shokyra, G. Ya. (2012). Correction priority early signs in constructing recognition systems. Naukovyi visnyk NLTU Ukrainy, 22.7, 344–350.
5. Jensen, R., Shen, Q. (2004). Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches. IEEE Transactions on Knowledge and Data Engineering, 16 (12), 1457–1471. doi: https://doi.org/10.1109/tkde.2004.96
6. Jain, A. K., Duin, P. W., Mao, J. (2000). Statistical pattern recognition: a review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22 (1), 4–37. doi: https://doi.org/10.1109/34.824819
7. Lavrakas, P. (2008). Encyclopedia of survey research methods. Sage Publications. doi: https://doi.org/10.4135/9781412963947
8. Dopico, J. R. R., Dorado, J., Pazos, A. (Eds.) (2009). Encyclopedia of artificial intelligence. IGI Global. doi: https://doi.org/10.4018/978-1-59904-849-9
9. Everitt, B. S., Landau, S., Leese, M., Stahl, D. (2011). Cluster Analysis. John Wiley & Sons, Ltd. doi: https://doi.org/10.1002/9780470977811
10. Ivakhnenko, A. H., Koppa, Yu. V. (1974). Vybir ansambliu oznak i syntez bahatoriadnoho pertseptrona za oznakamy samoorhanizatsiyi. Avtomatyka, 2, 41–53.

11. Ivahnenko, A. G., Koppa, Yu. V., Timchenko, I. K., Ivahnenko, N. A. (1980). Svyaz' teorii samoorganizatsii matematicheskih modeley na EVM i teorii raspoznavaniya obrazov. Avtomatika, 6, 3–13.
12. Ivahnenko, A. G. (1981). Induktivnyy metod samoorganizatsii modeley slozhnyh sistem. Kyiv: Naukova dumka, 296.
13. Gabor, D. (1971). Cybernetics and the Future of our Industrial Civilization. J. of Cybernetics, 1, 1–4.
14. Gabor, D. (1972). Perspektivy planirovaniya. Avtomatika, 2, 16–22.
15. Ivahnenko, A. G. (1989). Metod posledovatel'nogo oprobovaniya (perebora) klasterizatsiy-kandidatov po kriteriyam differentsial'nogo tipa. Raspoznavanie, klassifikatsiya, prognoz. Matematicheskie metody i ih primenenie, 2, 126–158.
16. Madala, H. R., Ivakhnenko, A. G. (1994). Inductive learning algorithms for complex systems modeling. CRC Press, 380. doi: https://doi.org/10.1201/9781351073493
17. Wójcik, W., Osypenko, V., Lytvynenko, V. (2013). The use of inductive clustering algorithms for forming expert groups in large-scale innovation projects. Elektronika: konstrukcje, technologie, zastosowania, 54 (8), 45–48. Available at: https://yadda.icm.edu.pl/baztech/element/bwmeta1.element.baztech-e864befd-7a77-411b-9ed7-44cb3446b06e
18. Babichev, S., Lytvynenko, V., Osypenko, V. (2017). Implementation of the objective clustering inductive technology based on DBSCAN clustering algorithm. 2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). doi: https://doi.org/10.1109/stc-csit.2017.8098832

The paper presents a new powerful technique to linearize the quadratic assignment problem. There are so many techniques available in the literature that are used to linearize the quadratic assignment problem. In all these linear formulations, both the number of variables and the linear constraints significantly increase. The quadratic assignment problem (QAP) is a well-known problem whereby a set of facilities are allocated to a set of locations in such a way that the cost is a function of the distance and flow between the facilities. In this problem, the costs are associated with a facility being placed at a certain location. The objective is to minimize the assignment of each facility to a location. There are three main categories of methods for solving the quadratic assignment problem. These categories are heuristics, bounding techniques and exact algorithms. Heuristics quickly give near-optimal solutions to the quadratic assignment problem. The five main types of heuristics are construction methods, limited enumeration methods, improvement methods, simulated annealing techniques and genetic algorithms. For every formulated QAP, a lower bound can be calculated. We have Gilmore-Lawler bounds, eigenvalue related bounds and bounds based on reformulations as bounding techniques. There are four main classes of methods for solving the quadratic assignment problem exactly, which are dynamic programming, cutting plane techniques, branch and bound procedures and hybrids of the last two. The QAP has application in computer backboard wiring, hospital layout, dartboard design, typewriter keyboard design, production process, scheduling, etc. The technique proposed in this paper has the strength that the number of linear constraints increases by only one after the linearization process

Keywords: quadratic assignment problem, Koopmans and Beckmann formulation, linear binary form

UDC 519
DOI: 10.15587/1729-4061.2021.225311

DEVELOPMENT OF A METHOD TO LINEARIZE THE QUADRATIC ASSIGNMENT PROBLEM

Elias Munapo
PhD, Professor of Operations Research
Department of Statistics and Operations Research
School of Economics and Decision Sciences
North West University
Mmabatho Unit 5, Mahikeng, Mafikeng, South Africa, 2790
E-mail: emunapo@gmail.com

Received date 11.02.2021
Accepted date 13.04.2021
Published date 30.04.2021

How to Cite: *Munapo, E. (2021). Development of a method to linearize the quadratic assignment problem. Eastern-European Journal of Enterprise Technologies, 2 (4 (110)), 54–61. doi: https://doi.org/10.15587/1729-4061.2021.225311*

1. Introduction

The quadratic assignment problem (QAP) is a well-known problem and this is a problem whereby a set of facilities are allocated to a set of locations in such a way that the cost is a function of the distance and flow between the facilities. In this problem, the costs are associated with a facility being placed at a certain location. The objective is to minimize the assignment of each facility to a location as given in [1, 2].

The QAP has application in wiring a computer backboard, in designing a hospital layout and in the dartboard design whereby in the game of darts points are scored by hitting specific marked areas of the board. The QAP is also used in the keyboard design of a typewriter, production process and scheduling.

2. Literature review and problem statement

In the paper [1], the QAP was linearized and the numbers of constraints and variables were kept to a minimal level. This linear formulation is based on the original nonlinear