http://doi.org/10.35784/iapgos.2700

GRANULAR REPRESENTATION OF THE INFORMATION POTENTIAL OF VARIABLES – APPLICATION EXAMPLE

Adam Kiersztyn[1], Agnieszka Gandzel[2], Maciej Celiński[2], Leopold Koczan[2]

[1]Lublin University of Technology, Department of Computer Science, Lublin, Poland, [2]Lublin University of Technology, Faculty od Technology Fundamentals, Lublin, Poland

Abstract. *With the introduction to the science paradigm of Granular Computing, in particular, information granules, the way of thinking about data has changed gradually. Both specialists and scientists stopped focusing on the single data records themselves, but began to look at the analyzed data in a broader context, closer to the way people think. This kind of knowledge representation is expressed, in particular, in approaches based on linguistic modelling or fuzzy techniques such as fuzzy clustering. Therefore, especially important from the point of view of the methodology of data research, is an attempt to understand their potential as information granules. In this study, we will present special cases of using the innovative method of representing the information potential of variables with the use of information granules. In a series of numerical experiments based on both artificially generated data and ecological data on changes in bird arrival dates in the context of climate change, we demonstrate the effectiveness of the proposed approach using classic, not fuzzy measures building information granules.*

Keywords: granular computing, information granules, knowledge representation, fuzzy clustering, ecological data

ZIARNISTA REPREZENTACJA POTENCJAŁU INFORMACYJNEGO ZMIENNYCH – PRZYKŁAD ZASTOSOWANIA

Streszczenie*. Wraz z wprowadzeniem do nauki paradygmatu obliczeń ziarnistych, w szczególności ziaren informacji, sposób myślenia o danych stopniowo się zmieniał. Zarówno specjaliści, jak i naukowcy przestali skupiać się na samych rekordach pojedynczych danych, ale zaczęli patrzeć na analizowane dane w szerszym kontekście, bliższym ludzkiemu myśleniu. Ten rodzaj reprezentacji wiedzy wyraża się w szczególności w podejściach opartych na modelowaniu językowym lub technikach rozmytych, takich jak klasteryzacja rozmyta. Dlatego szczególnie ważna z punktu widzenia metodologii badania danych jest próba zrozumienia ich potencjału jako ziaren informacji. W niniejszym opracowaniu przedstawimy szczególne przypadki wykorzystania innowacyjnej metody reprezentacji potencjału informacyjnego zmiennych za pomocą ziaren informacji. W serii eksperymentów numerycznych opartych zarówno na danych generowanych sztucznie, jak i danych ekologicznych dotyczących zmian dat przylotów ptaków w kontekście zmian klimatycznych, demonstrujemy skuteczność proponowanego podejścia przy użyciu klasycznych, a nie rozmytych miar budujących ziarna informacji.*

Słowa kluczowe: obliczenia ziarniste, ziarna informacji, reprezentacja wiedzy, grupowanie rozmyte, dane ekologiczne

Introduction

The problem of the correct selection of the model form (its type and recipe) is still an open problem, widely commented on in the world of science [2, 14, 19, 20]. Well-known approaches with an established reputation [6, 7] do not always perform well in comparison with modern machine learning techniques [16, 22]. One of the key issues, apart from the choice of the model form, is the proper selection of explanatory variables for the model. This issue has a huge impact on the entire course of building the model. There are a number of recognized techniques for selecting variables for the model [5, 8]. This problem is considered in many fields of science, such as ecology [9, 25], biology [15] and economics and sociology [1, 23]. However, most of these approaches propose specific, dedicated methods of selecting variables for a specific problem and model.

The main goal of this work is to design a method that enables the storage of information about the potential of individual variables included in a larger set of explanatory variables for the model. Therefore, the key challenge is to develop a method that stores as much information as possible in a readable manner. Moreover, it is assumed that this method is universal and independent of the analyzed issue. To solve this problem, we use an innovative concept of information pellets. The idea of building information pellets was born at the end of the last century, but in recent years it has been experiencing a renaissance [3, 4, 11–13, 18]. The results contained in the study are a crisp counterpart to the fuzzy theory presented in [13].

The work is organized as follows. The second part describes the proposed solutions. The third part presents two examples of the application of the proposed innovative solution. The final part contains conclusions and directions for future work.

1. Description of information potential granules

With a the set of $\boldsymbol{N}$ variables $\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}, \dots, \boldsymbol{X_N}$, each of which can be used as a potential explanatory variable in the model describing the selected variable, the potential usefulness of each of these variables should be determined. In the classical theory of model construction, a number of methods are used to select explanatory variables for the model. Our goal is to introduce a new representation of knowledge about the information potential of individual variables in an innovative way as information granules represented by numerical vectors.

In general, the information potential granule of each variable is a vector of the form:

$$[\omega_1 \colon \phi_1(X_1, X_2, X_3, \dots, X_N); \dots; \omega_k \colon \phi_k(X_1, X_2, X_3, \dots, X_N)] \quad (1)$$

where $\omega_i, 1 \leq i \leq k$ is the number of variables meeting the variable dependency criterion, while

$$\phi_i\ (X_1, X_2, X_3, \dots, X_N\ ), 1 \leq i \leq k \quad (2)$$

returns the numbers of variables for which the variable dependency criterion is met. The value of k represents the number of methods examining the relationship between the variables. Here, a separator ":" increases the transparency of the record by separating the number of result variables from their numbers. The variable dependency criterion can be any rule describing the relationship between the analyzed variables.

The basic criterion that is usually considered when checking whether the variable X_i is suitable for describing the variable X_j is the correlation between them. It is obvious that this classic approach should also be included in the proposed method for which information representing knowledge about how many variables (among all variables considered) and with which variables it is strongly correlated for a given variable X_i. In addition to the classic correlation for the variables themselves, it is also advisable to verify the increments of which variables are correlated with each other.

In addition, depending on the specifics of the analyzed variables, it is possible to add further components to the granule of information potential of the analyzed variable. For example, when individual variables describe similar phenomena, it is possible to determine how many and which variables always assume smaller (correspondingly larger) values than the analyzed variable. In many practical aspects, when analyzing complex data, it is important to know how nominally the corresponding values of the compared variables differ. Therefore, it is reasonable to specify the number and indices of variables whose values are in the channel determined by the variable values increased and decreased by a given value.

With information potential granules it is possible to build a network of relationships between individual variables, with the strength of the relationship between two variables dependent on the number of granule components for which both variables are related.

The course of action of the proposed innovative method of building granules of variable information potential is presented in Fig. 1.

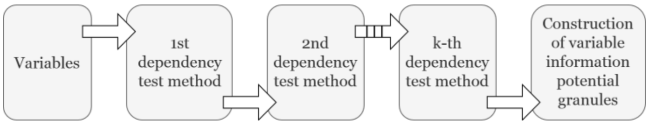
Variables
1st dependency test method
2nd dependency test method
k-th dependency test method
Construction of variable information potential granules

Fig. 1. Scheme of construction of information potential granules

2. Case study

The process of creating information granules in accordance with the described scheme of the construction of information granules is presented on examples of generated data and empirical data describing the arrival dates of migrating birds.

2.1. Generated data

Let us consider a set of 8 variables describing the values of a certain phenomenon changing over time. The values of the generated data analyzed are presented in the Table 1.

Table 1. Generated data

t	X1	X2	X3	X4	X5	X6	X7	X8
0	20	19	22	7	17	17	40	47
1	24	23	23	10	21	22	45	49
2	30	28	28	16	25	33	56	59
3	32	33	15	21	30	22	45	33
4	32	37	26	20	35	33	56	55
5	22	42	14	8	39	11	34	31
6	28	47	26	15	44	29	52	55
7	34	52	12	21	48	21	44	27
8	19	57	25	7	52	19	42	53
9	21	62	11	9	57	7	30	25
10	33	67	14	22	61	22	45	31
11	17	72	11	3	65	3	26	25
12	32	77	24	19	69	31	54	51
13	23	82	21	12	74	19	42	45
14	15	87	16	4	79	6	29	35
15	28	92	27	14	84	30	53	57
16	29	96	27	18	89	31	54	57
17	26	100	16	15	93	17	40	35
18	20	105	14	8	98	9	32	31
19	23	110	12	11	103	10	33	27
20	19	115	24	7	107	18	41	51

Now, let us consider the first variable i.e., X_1. The variables X_4, X_6 and X_7 are strongly correlated (Pearson's linear correlation coefficient values greater than 0.75) with this variable (see Fig. 2).

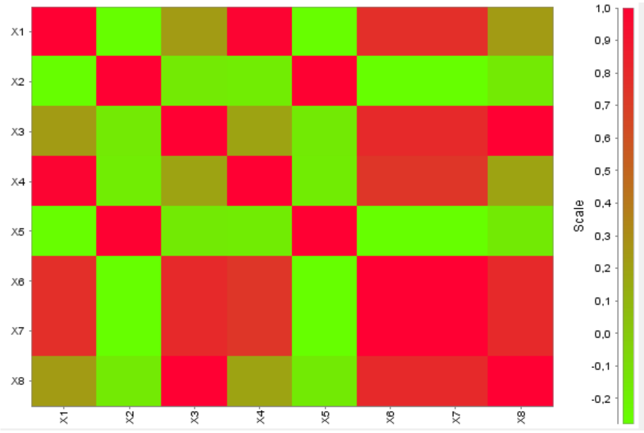

Fig. 2. The Pearson's linear correlation coefficient between variables X1 – X8

Figure 2 shows a heat map of the correlation between the different variables. As can be seen, according to the scale described in the figure, there are both strong relationships (red) and no correlation (green) between the analyzed variables.

In addition, if we determine the increments of individual variables and calculate the correlation of variable increments, we will notice that the same variables are strongly correlated with the X_1 variable, except that the values of correlation coefficients are slightly different from the original ones.

One of the easiest ways to determine linguistic descriptors that describe the behavior of variable values at different times is to use intervals based on basic statistical measures such as mean ($\bar{X}$) and standard deviation (s), or alternatively the median and quarter deviation. Examples of ranges of values and identifiers of individual descriptors with their descriptions are presented in Table 2.

Table 2. Examples of linguistic descriptors

Descriptor	Identifier	Range of values
Significantly less than typical value	-2	$(-\infty; \bar{X} - 1.5s]$
Slightly lower than typical value	-1	$(\bar{X} - 1.5s; \bar{X} - 0.5s]$
Typical value	0	$(\bar{X} - 0.5s; \bar{X} + 0.5s)$
Slightly larger than typical value	1	$[\bar{X} + 0.5s; \bar{X} + 1.5s)$
Significantly larger than typical value	2	$[\bar{X} + 1.5s; \infty)$

In the analyzed example, only one variable X_4 has values always smaller than the variable X_1, while X_7 takes higher values in all cases. In addition, the variables that differ (at any time) from the variable X_1 by at most 15 are variables X_4 and X_6.

On the individual components of the vector describing the information potential we have: linear Pearson correlation of individual variables, linear Pearson correlation of variable increments, linear Pearson correlation of linguistic descriptors. In all these cases, the numbers of the variables for which the coefficient exceeds the value of 0.75 are listed after the colon. Then, the numbers of variables for which values are always smaller than the considered variable (fourth coordinate of the vector), numbers of variables with values always greater than the variable under consideration (fifth coordinate) are listed, and finally the numbers of variables differing at each position by at least 15.

Therefore, based on the above considerations, for the variable X_1, the information granules describing its potential, according to formula (1), takes the form:

$$X_1 = [3:4,6,7; 3:4,6,7; 1:4; 1:4; 1:7; 2:4,6].$$

After conducting analogous considerations for the remaining variables, we obtain the following granules of variable information potential:

$$X_2 = [1:5; 0; 1:5; 2:4,5; 0; 1:5];$$
$$X_3 = [3:6,7,8; 3:6,7,8; 3:6,7,8; 0; 2:7,8; 1:6];$$
$$X_4 = [1:1; 1:1; 1:1; 0; 5:1,2,5,7,8; 1:1];$$
$$X_5 = [1:2; 0; 1:2; 1:4; 1:2; 1:2];$$
$$X_6 = [4:1,3,7,8; 4:1,3,7,8; 3:3,7,8; 0; 2:7,8; 2:1,3];$$
$$X_7 = [4:1,3,6,8; 4:1,3,6,8; 3:3,6,8; 4:1,3,4,6; 0; 0];$$
$$X_8 = [3:3,6,7; 3:3,6,7; 3:4,6,7; 3:3,4,6; 0; 0].$$

As it can be seen, the variable X_2 has a much lower potential than the variable number 1, which is confirmed by the smaller number of variables entering the relation with variable 2. The variables that seem to have the greatest information potential are 6 and 7. It remains an open question to decide which of these variables is more useful. Partial answer to this question comes from the analysis of the network of relationships between the variables.

The network of relationships for variables from the case considereds would look like the one presented in Fig. 3.

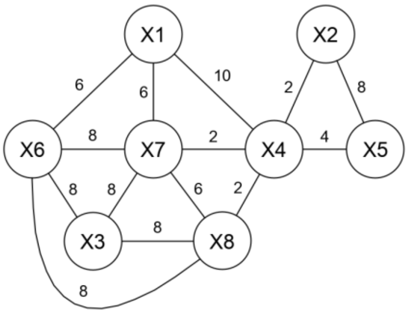

Fig. 3. Dependence network based on granules of variable information potential

Analyzing the network of connections from Fig. 3, we notice that there are the most links between variables X_1 and X_4, while the variable X_5 is in fact directly related only to the variable X_2.

The variables with the greatest total potential are X_6 and X_7 . However, the variable X_7 seems to be the variable with the greatest potential for use in building models that reflect the behavior of other variables, although this variable is not suitable for modeling variables X_2 and X_5.

2.2. Analysis of birds' first arrival dates

Modern climate changes affect many animal species, including birds. One of the issues widely discussed in the scientific literature related to the birds' response to climate change are shifts in the dates of spring arrivals and their ecological consequences [24]. However, to analyze such phenomena in historical terms, an appropriate indicator is needed. For a migratory bird species which is absent from a location for some of the year, the simplest phenological measure is the earliest report of its return, i.e., the first arrival date (FAD) [17]. Recording FADs has long appealed to human nature and it has been possible to extract historical data providing information on phenological changes expanding back two centuries [17]. Recently, annual compilations of local or regional bird records have begun to include this information as standard [17]. In this paper we used the FADs of 79 species of birds collected in northern Poland (the North Podlasie Lowland: 21°51′–23°57′ E; 52°17′–53°54′ N) in the years 1996-2016 [10, 21].

We will analyze 26 variables describing the arrival dates of birds (26 different species) [10] in close migration in the years 1996–2016. The names of individual analyzed species along with the assigned variable numbers are presented in Table 3**.**

First arrival dates for individual species are presented in table 4.

As in the case of generated data, we assume that the criteria for relationships between variables can be a correlation calculated directly for variables, correlation of annual increments, correlation for descriptors based on the mean (according to the formulas presented in Table 2), correlation of arrival weeks (instead of the day of the year, the weeks of arrival are considered and the correlation is calculated for them), incoming species always before the analyzed species, arriving always after the analyzed species, species for which the difference in arrival times does not exceed 3 weeks. All correlation coefficients with the acceptance threshold as significant equal to 0.7.

Then, following similar considerations as in the previous point, we can obtain the list of individual components of the information granules, based on formula (1), which is presented in table 5.

It shoud be noted that some variables (bird species) seem to have greater potential to describe other variables at first glance. In order to increase the transparency of the analysis of the obtained results, a heat map (see Fig. 4) was developed, which clearly shows the relationships between individual variables (bird species).

Table 3. Species names of closely migrating birds that have been analyzed

Variable index	English name	Latin name
1	Eurasian bittern	Botaurus stellaris
2	Marsh Harrier	Circus aeruginosus
3	Lapwing	Vanellus vanellus
4	Wood Pigeon	Columba palumbus
5	Eurasian blackcap	Sylvia atricapilla
6	Black Redstart	Phoenicurus ochruros
7	Spotted Crake	Porzana porzana
8	Common Redshank	Tringa totanus
9	Common snipe	Gallinago gallinago
10	European serin	Serinus serinus
11	Woodlark	Lullula arborea
12	Coot	Fulica atra
13	Great crested grebe	Podiceps cristatus
14	Red-necked grebe	Podiceps grisegena
15	Common Chiffchaff	Phylloscopus collybita
16	White wagtail	Motacilla alba
17	Dunnock	Prunella modularis
18	Reed Bunting	Schoeniclus schoeniclus
19	Eurasian penduline-tit	Remiz pendulinus
20	Eurasian Skylark	Alauda arvensis
21	Eurasian Woodcock	Scolopax rusticola
22	Black-headed gull	Chroicocephalus ridibundus
23	Song Thrush	Turdus philomelos
24	Meadow Pipit	Anthus pratensis
25	Western Water Rail	Rallus aquaticus
26	Common Crane	Grus grus

Table 4. First arrival day for individual species in 1996–2016

Year	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23	X24	X25	X26
1996	108	88	45	62	115	56	114	70	78	87	61	85	80	84	96	67	97	61	118	44	70	49	66	61	89	59
1997	85	90	53	61	119	83	105	87	72	114	60	91	92	122	112	70	111	56	122	43	89	57	55	53	92	59
1998	82	57	45	84	108	90	101	79	85	100	61	51	84	91	95	86	94	53	95	47	79	54	67	74	96	53
1999	91	82	63	80	114	82	116	79	79	88	63	65	79	88	87	75	82	63	88	59	90	63	61	78	116	65
2000	85	68	44	71	109	75	101	73	84	108	73	68	68	84	98	73	85	47	84	42	89	63	84	73	98	57
2001	100	76	40	91	98	82	93	75	69	94	71	73	76	91	90	69	92	48	91	40	76	69	69	70	94	59
2002	75	71	51	75	107	84	105	76	72	79	69	70	80	88	92	41	85	47	80	34	49	48	71	72	111	56
2003	106	81	52	71	109	88	107	82	71	82	70	74	80	103	102	75	96	70	95	67	51	71	70	71	103	66
2004	91	79	73	76	99	89	107	77	73	101	80	77	79	91	89	77	82	77	87	67	79	72	73	73	97	73
2005	86	78	73	81	98	84	100	83	83	93	69	83	85	100	91	78	87	72	90	74	91	76	75	72	102	72
2006	94	84	81	83	101	83	105	84	82	93	77	85	89	97	91	83	92	82	95	64	90	83	77	83	92	78
2007	80	66	60	51	101	62	101	66	66	86	61	67	84	93	85	66	83	64	88	48	72	64	62	66	72	61
2008	75	74	55	53	100	60	104	69	67	79	52	60	80	94	87	69	86	58	90	48	60	66	63	64	81	68
2009	75	78	58	69	96	89	97	71	73	86	69	58	81	92	91	73	81	70	92	47	79	58	60	70	97	69
2010	80	79	59	66	94	79	89	82	79	86	71	75	65	88	85	56	82	75	87	45	83	62	64	66	107	55
2011	75	79	59	66	94	79	89	82	79	86	71	75	65	88	85	56	82	70	87	45	83	62	60	66	107	55
2012	83	68	56	72	98	73	81	75	72	87	69	71	59	77	82	76	85	64	79	52	74	56	63	60	98	56
2013	100	96	65	87	96	89	104	83	96	103	95	81	89	104	99	66	97	65	102	39	100	49	63	96	107	62
2014	67	63	46	69	67	88	87	65	67	82	63	60	85	82	104	48	68	45	58	46	72	34	100	58	82	43
2015	76	80	54	53	109	80	94	74	68	86	67	79	74	79	85	53	83	53	88	53	64	62	67	67	76	53
2016	85	79	56	71	101	83	101	79	79	91	69	74	80	91	91	71	85	62	90	47	79	62	67	71	97	57

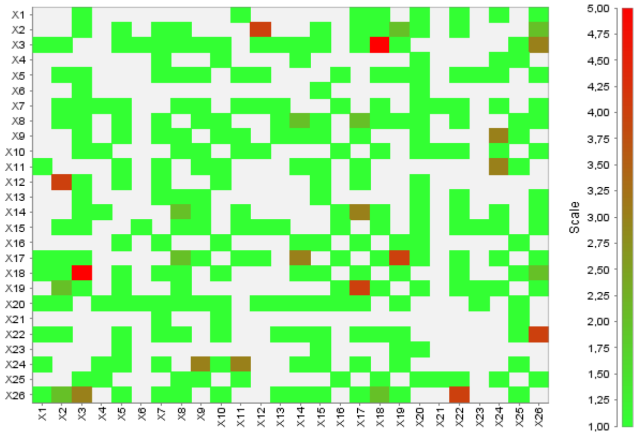

Fig. 4. Heat map presenting the network of connections between variables (species)

Analyzing the results presented in Fig. 4, it should be noted that among the considered species of birds are those that potentially very well describe the arrival times of other species. The pairs of variables between which there are no relations described above are marked in white. For the variables between which there are relationships, the colour scale presented in Fig. 4 was used. The red colour marks the variables between which there is the largest number of the relationships under consideration. The species that has the greatest potential for explaining other species is "*Lapwing*" (saved as variable X_3). This species has an information potential of 23 and is associated with 17 other species.

In addition, "*Lapwing*" is strongly associated with "*Reed bunting*" (variable X_{18}). The information potential between these species is 5, i.e. for five different methods of testing relationships between variables there is a significant relationship between these variables (species). Other potentially good explanatory variables (species that characterize other species well) are the variables X_{17}, X_{18} and X_{26} (i.e. "*Dunnock*", "*Reed bunting*" and "*Common crane*"). In turn, variables with the lowest information potential are the variables X_4, X_6, X_{21} and X_{23} (i.e. "*Wood pigeon*", "*Black redstart*", "*Eurasian woodcock*" and "*Song thrush*"). These species do not have a lot of common features with other species, and therefore they will relatively rarely be used in the construction of behavioral models of other species.

3. Conclusion and future works

In this study, we have considered an emerging paradigm of information granules to establish the granular information potential of the variables. We have thoroughly examined the proposed method using the set of information about migrating birds to find the dependencies between the ways various species behave. The results of numerical experiments have shown the potential hidden in the method. Future directions of the studies may cover, among others, an application of fuzzy set-based techniques to build information potential granules as well as other than fuzzy works with uncertainty in the data. Moreover, it would be worth examining the method utilizing datasets containing more complex information, e.g., that coming from logistics, weather forecasting, or financial datasets containing time series.

Table 5. The components of information granules describing the potential of individual

Variable index	Correlation	Increment correlation	Descriptor correlation	Week correlation	Always before	Always after	Channel
1	0	0	1:17	0	7:3,11,18,20,22,24,26	0	0
2	1:12	4:8,12,19,26	1:19	1:12	5:3,18,20,22,26	4:5,7,15,17	1:12
3	2:18,26	2:18,26	1:18	1:18	0	15:1,2,5,6,7,8,9,10,12,13,14,15, 17,19,25	2:18,26
4	0	0	0	0	1:20	4:7,10,14,25	1:24
5	1:19	1:23	0	0	12:2,3,8,9,11,12,16,18,20,22,24,26	0	1:7
6	0	0	0	0	2:3,20	1:15	0
7	0	0	0	0	15:2,3,4,8,9,11,12,13,16,18,20,21,22,24,26	0	1:5
8	0	4:2,14,17,19	0	0	5:3,18,20,22,26	7:5,7,10,14,15,17,25	2:9,13
9	1:24	2:11,24	0	0	2:3,20	7:5,7,10,14,15,17,25	3:8,13,24
10	0	1:21	0	0	12:3,4,8,9,11,12,16,18, 20,22,24,26	0	0
11	1:24	2:9,24	0	0	0	8:1,5,7,10,14,15,17,25	1:24
12	1:2	1:2	0	1:2	2:3,20	5:5,7,10,15,17	1:2
13	0	0	0	0	4:3,20,22,26	2:7,15	2:8,9
14	1:17	2:8,17	0	0	11:3,4,8,9,11,16,18,20, 22,24,26	0	1:17
15	0	0	0	0	15:2,3,6,8,9,11,12,13,16,18,20,22,23,24,26	0	0
16	0	0	0	0	1:20	8:5,7,10,14,15,17,19, 25	0
17	2:14,19	3:8,14,19	1:1	1:19	12:2,3,8,9,11,12,16,18,20,22,24,26	0	2:14,19
18	1:3	2:3,26	1:3	1:3	0	11:1,2,5,7,8,10,14,15, 17,19,25	3:3, 22, 26
19	2:5,17	4:2,8,17,23	1:2	1:17	7:3,16,18,20,22,24,26	0	1:17
20	0	0	0	0	0	18:1,2,4,5,6,7,8,9,10, 12,13,14,15,16,17,19,23,25	0
21	0	1:10	0	0	0	2:7,25	0
22	1:26	0	1:26	1:26	0	12:1,2,5,7,8,10,13,14, 15,17,19,25	2:18, 26
23	0	2:5,19	0	0	1:20	1:15	0
24	2:9,11	2:9,11	0	0	0	9:1,5,7,10,14,15,17,19,25	3:4, 9,11
25	0	0	0	0	12:3,4,8,9,11,16,18,20,21,22,24,26	0	0
26	2:3,22	3:2,3,18	1:22	1:22	0	12:1,2,5,7,8,10,13,14,15,17,19,25	3:3, 18, 22

References

[1] Altonji J. G., Elder T. E., Taber C. R.: Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. Journal of Political Economy 113(1), 2005, 151–184 [http://doi.org/10.1086/426036].

[2] Barbieri M. M., Berger J. O.: Optimal predictive model selection. Ann. Statist. 32(3), 2004, 870–897 [http://doi.org/10.1214/009053604000000238].

[3] Bargiela A., Pedrycz W.: Human-centric information processing through granular modelling. Springer Science & Business Media 182, 2009 [http://doi.org/10.1007/978-3-540-92916-1].

[4] Bargiela A., Pedrycz W.: Granular computing. In: Handbook on Computational Intelligence. World Scientific, 2016 [http://doi.org/10.1142/9789814675017_0002].

[5] Bursac Z., Gauss, C. H., Williams D. K., Hosmer D. W.: Purposeful selection of variables in logistic regression. Source Code for Biology and Medicine 3(1), 2008, 17 [http://doi.org/10.1186/1751-0473-3-17].

[6] Gauch H.: Model selection and validation for yield trials with interaction. Biometrics 44(3), 1988, 705–715 [http://doi.org/10.2307/2531585].

[7] Geisser S., Eddy W. F.: A predictive approach to model selection. Journal of the American Statistical Association 74(365), 1979, 153–160 [http://doi.org/10.1080/01621459.1979.10481632].

[8] Genuer R., Poggi J. M., Tuleau-Malot C.: Variable selection using random forests. Pattern Recognition Letters 31(14), 2010, 2225–2236 [http://doi.org/10.1016/j.patrec.2010.03.014].

[9] Johnson J. B., Omland K. S.: Model selection in ecology and evolution. Trends in Ecology & Evolution 19(2), 2004, 101–108 [http://doi.org/10.1016/j.tree.2003.10.013].

[10] Kiersztyn A., Karczmarek P., Lopucki R., Pedrycz W., Al E., Kitowski I., Zbyryt A.: Data imputation in related time series using fuzzy set-based techniques. IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Glasgow 2020, 1–8.

[11] Kiersztyn A., Karczmarek P., Kiersztyn K., Pedrycz W.: Detection and Classification of Anomalies in Large Data Sets on the Basis of Information Granules. IEEE Transactions on Fuzzy Systems, 2021 [http://doi.org/10.1109/TFUZZ.2021.3076265].

[12] Kiersztyn A., Karczmarek P., Kiersztyn K., Pedrycz W.: The Concept of Detecting and Classifying Anomalies in Large Data Sets on a Basis of Information Granules. IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2020, 1–7.

[13] Kiersztyn A., Karczmarek P., Kiersztyn K., Łopucki R., Grzegórski S., Pedrycz W.: The Concept of Granular Representation of the Information Potential of Variables. 2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2021, 1–6.

[14] Laud P.W., Ibrahim J.G.: Predictive model selection. Journal of the Royal Statistical Society: Series B (Methodological) 57(1), 1995, 247–262 [http://doi.org/10.1111/j.2517-6161.1995.tb02028].

[15] Mac Nally R.: Regression and model-building in conservation biology, biogeography and ecology: the distinction between – and reconciliation of – "predictive" and "explanatory" models. Biodiversity & Conservation 9(5), 2000, 655–671 [http://doi.org/10.1023/A:1008985925162].

[16] Olivera A. R., Roesler V., Iochpe C., Schmidt M. I., Vigo A., Barreto S. M., Duncan B. B.: Comparison of machine-learning algorithms to build a predictive model for detecting undiagnosed diabetes-elsa-brasil: Accuracy study. Sao Paulo Medical Journal 135(3), 2017, 234–246 [http://doi.org/10.1590/1516-3180.2016.0309010217].

[17] Pearce-Higgins J. W., Green R. E.: Birds and climate change: Impacts and conservation responses. Cambridge University Press 2014.

[18] Pedrycz W.: Knowledge-based clustering: From data to information granules. John Wiley & Sons, 2005 [http://doi.org/10.5555/1044924].

[19] Piironen J., Vehtari A.: Projection predictive model selection for Gaussian processes. IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP), Salerno 2016, 1–6.

[20] Piironen J., Vehtari A.: Comparison of Bayesian predictive methods for model selection. Statistics and Computing 27(3), 2017, 711–735. [http://doi.org/10.1007/s11222-016-9649-y].

[21] ptop.org.pl (2016), (available: 01.10.2020).

[22] Schafer B. C., Wakabayashi K.: Machine learning predictive modelling high-level synthesis design space exploration. IET Computers & Digital Techniques 6(3), 2012, 153–159 [http://doi.org/10.1049/iet-cdt.2011.0115].

[23] Smith A., Naik P. A., Tsai C. L.: Markov-switching model selection using Kullback-Leibler divergence. Journal of Econometrics 134(2), 2006, 553–577 [http://doi.org/10.1016/j.jeconom.2005.07.005].

[24] Stephens P. A., Mason L. R., Green R. E., Gregory R. D., Sauer J. R., Alison J., Aunins A., Brotons L., Butchart S. H., Campedelli T., et al.: Consistent response of bird populations to climate change on two continents. Science 352(6281), 2016, 84–87 [http://doi.org/10.1126/science.aac4858].

[25] Symonds M. R., Moussalli A.: A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. Behavioral Ecology and Sociobiology 65(1), 2011, 13–21 [http://doi.org/10.1007/s00265-010-1037-6].

Ph.D. Adam Kiersztyn
e-mail: a.kiersztyn@pollub.pl

He received the Ph.D. degree in mathematics from Faculty of Mathematics, Physics, and Computer Science, Maria Curie-Skłodowska University, Lublin, Poland, in 2012. He is currently an assistant professor with the Department of Computer Science, Lublin University of Technology, Lublin, Poland. His current research interests include fuzzy measures, data mining, data exploration, quantitative methods, and decision-making theory.

http://orcid.org/0000-0001-5222-8101

Ph.D. Agnieszka Gandzel
e-mail: a.gandzel@pollub.pl

In 2015, she received her Ph.D. in pedagogy from the Faculty of Social Sciences at the John Paul II Catholic University of Lublin, Poland. She is currently an assistant professor at the Faculty of Technology Fundamentals, Lublin University of Technology, Poland. Her research area is the application of new technologies in education.

http://orcid.org/0000-0002-7887-8636

M.Sc. Maciej Celiński
e-mail: m.celinski@pollub.pl

He received the degree in M.Sc. informatics, faculty of exact sciences at the John Paul II Catholic University of Lublin, Poland. He is currently an assistant at the Faculty of Technology Fundamentals, Lublin University of Technology, Lublin. His current research interests ICT and new technology
in education.

http://orcid.org/0000-0001-8412-207X

D.Sc. Leopold Koczan
e-mail: l.koczan@pollub.pl

He received his Ph.D. in mathematics in 1978 and his D.Sc. in 1989. He was Head of the Department of Applied Mathematics at the Faculty of Mechanical Engineering at the Lublin University of Technology, and also worked at the Faculty of Technology Fundamentals in the Department of Applied Mathematics.

http://orcid.org/0000-0002-7775-1836

otrzymano/received: 05.07.2021 *przyjęto do druku/accepted: 15.09.2021*