THE FORMATION OF A SET OF INFORMATIVE FEATURES BASED ON THE FUNCTIONAL RELATIONSHIPS BETWEEN THE DATA STRUCTURE FIELD OBSERVATIONS

Artemenko M.V., Kalugina N.M., Dobrovolsky I.I.

South-West State University, Kursk,
e-mail: artem1962@mail.ru, nat91art@mail.ru, iidobrovolskii@gmail.com

The methods of forming the set of informative features – tuple linguistic variables to solve diagnostic tasks in a decision support system diagnostic decision making in medicine. It is proposed to use the parameters of the approximating polynomials, algebraic and logical functions, correlations and criteria exploration of clustering for the formation of a variety of signs and calculation of informativeness based on rank sorting. Formulate the paradigm of the formation of each alternative node of the hierarchical decision tree differential diagnostic private sets of informative indicators.

Keywords: informativeness, approximately polynomial, differential diagnosis, method of hierarchies, tuple linguistic variable

Modern medical service of the population based on the information and computer technologies to support various stages of treatment and diagnostic process [6]. The development of the theoretical basis and software tools of artificial intelligence for solving tasks of classification and pattern recognition, forecasting led to the creation of various specialized automated systems of support of acceptance diagnostic solutions (ASSADS) for the tasks of clinical medicine and training of health workers [1, 2, 6, 16].

Design specialized ASSADS in medicine is based on the formation of adequate and effective knowledge base on the basis of decisive diagnostic rules synthesized and tested on clinically confirmed material, each element of which is characterized by a certain multiple of the recorded monitored and managed characteristics of the biological object or process. The problem of forming the set of informative features is important because the quality of its resolution depends on the efficiency of further diagnosis, as with the use and without the use of automated ASSADS.

From a medical point of view, the formation of extensive information, many signs bear semantic load as the formation of the tuple linguistic variables for the symptoms of a particular disease or condition of the body.

Feature build ASSADS for clinical medicine is the use in real conditions small amounts of training and examination (control) of samples of research results state of the biological object or process. Necessary and sufficient conditions imposed on the volume of the investigated material from the point of view of classical evidence-based medicine almost unrealizable in terms of the analysis of open systems (which are objects), vagueness and inaccuracies of recorded data in conditions of uncertainty. In addition, the same system of signs may have an acceptable informative for solving a recognition task and a completely unsuitable for another [13].

Formation tuple linguistic variables (many informative features) is a subject of many studies, fundamental of which are the work of G.S. Foreheads (e.g. [10]). Consider a number of methods of forming a tuple (as previously studied and proposed by the authors) based on the methodologies: the analytic hierarchy process (ordering is based on has go obtained grades – weights), regression analysis and self-organization of structural-parametric identification of mathematical models of the method of group accounting of arguments, or logical functions (identified, for example, logic algorithms, artificial neural networks [5]).

In the beginning of the study the characteristics set non-formalized way, with the help of experts (the Delphi technique or the fuzzy Delphi method) [7], recommended for the analysis of biomedical information due to its registration) or forcibly, taking into account the personal experience and knowledge of the researcher and analysis of specialized literature.

Then, it is proposed to apply the following, proven in practice methods and algorithms [8, 9]: Full – exhaustive of various combinations of signs to achieve acceptable diagnostic effect, Add – sequentially adds features; Del – sequential elimination of symptoms until disappearance of the previous diagnostic effect; AddDel – the simultaneous execution of the procedures of the algorithms Add and Del; Prob – for each attribute are determined by weight and then applied the procedure of

the above algorithms; fractal analysis applied to the tensor data (e.g., diagnosis of Parkinson's disease); Grad is the same as algorithm AddDel, but the inclusion and exclusion of indicators in the resulting lot is not "one", and "complex".

(Note that as features are directly measured and latent or integral, as the latter can be used indicators of system organization whose application is considered in [3, 4]).

These algorithms analyze the characteristics of the data structure, which is suggested to use the coefficients of pair correlation and/or the distance to the cluster centers. In this case, it is recommended to apply criteria – quality indicators [7]: Given index, indices of density, total Giprobum, the index of the Davis – Bouldin. I.e., a small volume of the sample applied these algorithms and indicators of the quality of a certain value, the generated sets of linguistic variables consisting of specific symptoms. In this case, the researcher specifies the "freedom of choice decision-making" – the number of sets from which to exam the sample according to the external criteria retained are the most informative.

If the implementation of exploratory cluster analysis is impossible, it is proposed that a simple and semantically transparent method in the final set of linguistic variables retained those characteristics that have the least correlation with the left and the highest with "discarded".

For deciding on the inclusion of symptom information, many are encouraged to use the methodology of decision making T.L. Saaty [14]. Create a matrix of preference of the elements of W, which elements to indices i and j differ by 9 degrees (the sign of i is preferable than attribute j): $w_{i,j} = 1$ – equal preference, $w_{i,j} = 2$ – the low degree of preference, $w_{i,j} = 3$ – medium preference, $w_{i,j} = 4$ – a preference above average, $w_{i,j} = 5$ – moderately strong preference, $w_{i,j} = 6$ – a strong preference, $w_{i,j} = 7$ – very strong (obvious) preference, $w_{i,j} = 8$ – a very, very strong preference, absolute preference, $w_{i,j} = 9$ – absolute preference.

Analysis of the matrix allows conversion of the matrix to group the signs by clusters of preference with the IJ-conversion. Is a permutation of the row I with row J in the matrix of modified preferences so that around the main diagonal of the clustered matrix elements with the highest values. The stop condition of the process of permutation acts achieve the minimum sum-of-products of the element values of the modified preference matrix W^* the distance of this element from the main diagonal according to the formula:

$$\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\left(w_{i,j}^{*}\cdot|j-i|\right)\right)\rightarrow \min, \qquad (1)$$

where $w_{i,j}^{*}=\begin{cases} w_{i,j}, & for \quad i<j; \\ 10-w_{i,j}, & for \quad i>j; \end{cases}$ N – the number of analysed characteristics before selection.

The degree of preference are proposed to determine by way of order signs on ranks of informativeness in descending order. The rank of informativeness metric for SPDR diagnostic character proposed to determine one way (or all – given the known algorithms of decision making on several alternative two).

Method 1. – By the maximum gradient of the functional differences (MGR) with or without taking into account latent integral indicator of systemic organization of functional States (proposed and approved by school A.V. Zavyalov – [3]);

Method 2. By analysing the structure and the parameters of the approximating polynomial Gabor [15].

Method 3. By analysing the structure and analysis of Boolean functions obtained by applying the algorithms and software logic, artificial neural networks [5].

Method 4. In terms of clustering quality [7].

In the first proposed method, for each alternative class is the matrix of pair connectivity (for example, the absolute value of the Pearson correlation coefficient) between variables, whose elements equal zero, if the calculated value is less than a certain threshold level. Classes for characteristics that are candidates for inclusion in the informative tuple linguistic variables are determined by the number of links – $Ks_i^{w_0}$ and $Ks_i^{w_1}$and calculated differences (gradients) $G_i=\left|Ks_i^{w_0}-Ks_i^{w_1}\right|$, for which the signs of i in descending order the Ks_i. For the ordered set of indicators are the ranks Rn_i by the formula:

$$Rn_i=\begin{cases} N, & for \quad i=1; \\ Rn_i-1, & for \quad (i\neq 1)\&(G_i\neq G_{i-1}); \\ Rn_i, & for \quad (i\neq 1)\&(G_i=G_{i-1}). \end{cases} \qquad (2)$$

The vector {*Rn*} is the matrix of preferences *W*, the values of the elements of which are calculated in accordance with the gradation proposed by T.L. Saaty (presented earlier) or cognitology or automatically – by the formula

$$w_{i,j} = \begin{cases} \left[\dfrac{9\cdot x + \max x - 9\cdot \min x}{\max x - \min x}\right], & for\ i < j \\ 1 - \dfrac{9\cdot x + \max x - 9\cdot \min x}{\max x - \min x}, & for\ i > j \end{cases}, \quad (3)$$

where $x = \left|Rn_i - Rn_j\right|$; $\max x = \max_i \left|Rn_i - Rn_j\right|$;

$\min x = \min_i \left|Rn_i - Rn_j\right|$; $w_{i,i} = 9$.

The second method of forming a matrix of preferences of information content of signs involves the use of the approximating polynomial Gabor – formula (4), since the increase in the degree of the polynomial the accuracy of the approximation, they approximated the function increases and then decreases – this allows you to apply a polynomial in the self-organizing algorithms of the group method of accounting arguments (GMDH) [11, 12]. Note that the GMDH allows handling samples of small volume and building the Gabor polynomial at the interpolation nodes, the number of which is smaller than the maximum degree of the polynomial.

$$Y(Z) = A_0 + \sum_{k=1}^{L}\left(A_k \cdot \prod_{i=1}^{N} z_i^{p_{i,k}}\right), \quad (4)$$

where $Z = \{z_1, z_2, \ldots, z_N\}$ – a lot of arguments; $Y(Z)$ is the response function (approximant); L is the number of terms in the polynomial; A_k, $p_{i,k}$ – the identified model parameter; N is the number of arguments

The information content of the indicator of the set {*X*} is proposed to define the following methods:

1 method – based nonlinear discriminant functions identified for class w_1 and w_0 ("ill" – "not ill", "condition 1" – "condition 2" – i.e., assumes a binary hierarchical decision tree). According to the recommendations of [6] for a class w_0 sets the value of the response function that lies in the range $(-1 \pm e)$ and having a uniform distribution $\left(e = \dfrac{1}{N_0 + N_1}\right)$, where N_0, N_1 – volume training samples for class w_0 and w_1, respectively). Similarly formed response for a class w_1 in the range $(1 \pm e)$ of the formula (4) and using the orthogonal algorithm GMDH is the structural-parametric identification of a polynomial (4).

Next, we determine the share of influence of each term in formula in each class:

$$V_k^{w_{0/1}} = \frac{A_k^{w_{0/1}} \cdot \overline{\prod_{i=1}^{N} x_i^{p_{i,k,w_{0/1}}}}}{\sum_{j=1}^{L}\left(A_j^{w_{0/1}} \cdot \overline{\prod_{i=1}^{N} x_i^{p_{i,j,w_{0/1}}}}\right)}, \quad (5)$$

where the operator $(\overline{ZZ})$ – represents the modal value of *ZZ*.

Then, for each argument included in the *k*-th term is calculated the weight of multiplicanda by the formula

$$M_{i,k}^{w_{0/1}} = \frac{\left|p_{i,k,w_{0/1}}\right| \cdot \left|\ln(\overline{x}_i)\right|}{\sum_j^N \left|p_{i,k,w_{0/1}}\right| \cdot \left|\ln\right|}. \quad (6)$$

In the end, determines the value of additive-multiplicative effect of indicator xi on the response function (according to the parameters of the discriminant approximant) for each alternative class, according to the formula

$$AM_{x_i}^{w_{0/1}} = 1 - \prod_{k=1}^{L}\left(1 - V_k^{w_{0/1}} \cdot M_{i,k}^{w_{0/1}}\right). \quad (7)$$

Introduces a relative error of "difference" $\varepsilon < 0{,}5$ (recommended of $0{,}01 \le \varepsilon < 0{,}1$) and recalculated the values of the multiplicative effects in $\left[AM_{x_i}^{w_{0/1}}, \varepsilon\right]$ by the formula (8):

$$\left[AM_{x_i}^{w_{0/1}}, \varepsilon\right] = \begin{cases} AM_{x_j}^{w_{0/1}}, & if\ (1-\varepsilon)\cdot AM_{x_j}^{w_{0/1}} < AM_{x_i}^{w_{0/1}} \le (1+\varepsilon)\cdot AM_{x_j}^{w_{0/1}}; \\ AM_{x_i}^{w_{0/1}}, & otherwise, \end{cases} \quad j = \overline{1,N},\ j \ne i. \quad (8)$$

Next, for each class (w_1 and w_0) signs (linguistic variables) out in descending order of values of $\left[AM_{x_i}^{w_{0/1}}, \varepsilon\right]$. In the end, are formed two-tuple of signs for classes: $\{X^{w_0}\}$ and $\{X^{w_1}\}$. According to the obtained tuples by applying the formula (2), replacing G_i for $\left[AM_{x_i}^{w_{0/1}}, \varepsilon\right]$ generated two sets of ranks Rn^{w_0} and Rn^{w_1}.

By Rn^{w_0} and Rn^{w_1} finalized many informative features according to a specific researcher volume $NI \le N$ consisting of elements $XI_j = \left(x_j / Inf\left(x_j\right)\right) / j = \overline{1,NI}$, which are imported from the original set $\{X\}$ according to Rn^{w_0} and Rn^{w_1} in descending order by serial connection in descending order of ranks. In case of alternative situations inclusions apply one of the following: «handcontol» (knowledge and experience of the researcher), Monte-Carlo, or by reducing the magnitude of ε, and repeat the procedure of ranking.

The information content of sign $Inf(x_j)$ proposes is determined by the formula

$$Inf\left(x_j\right) = \frac{\max\left(Rn_{x_j}^{w_0}, Rn_{x_j}^{w_1}\right)}{\max\limits_{j=\overline{1,NI}}\left(\max\left(Rn_{x_j}^{w_0}, Rn_{x_j}^{w_1}\right)\right)}, \quad (9)$$

where $Rn_{x_j}^{w_0}$, $Rn_{x_j}^{w_1}$ – value of rank metric x_j in w_0 and w_1, respectively.

2 method of forming the set of informative indicators, and the calculation of $Inf(x_j)$, based on preliminary identification of the approximating polynomial Gabor (4) for each indicator from the initial set $\{X\}$. In this case, the identification procedure is repeated N times for each class w_0 and w_1, sequentially forming the set $\{Z\} = \{X\} - x_j$ and responses $Y(Z) = x_j$.

As a result, generated many approximants for alternative classes:

$$\{App\}_{M_0}^{w_0} \quad \text{and} \quad \{App\}_{M_1}^{w_1}$$

$$(M_0 \le N,\ M_1 \le N,\ M_0 \ne 0,\ M_1 \ne 0).$$

It should be noted that approximate with values of coefficient of determination less than a certain researcher thresholds in further analysis is not involved. If the result of selection produced an empty lot approximants, it consistently returned approximant with the highest values of determination coefficients. The minimum amount many approximativeness "freedom of choice" (the recommended value of 3 to 7).

Next, for each alternative class formed matrix $(ApX)_{M_0,N}^{w_0}$ and $(ApX)_{M_1,N}^{w_1}$, the number of rows which are equal, respectively, M_0 and M_1, number of columns – number of indicators the set $\{X\}$, the value of the element matrices are calculated using formulas similar to (5)–(8). On the resulting matrices to form two vectors $(SApX)_N^{w_0}$ and $(SApX)_N^{w_1}$ (for each class), the values of which are calculated by formulas (10):

$$SApX_i^{w_0} = \max_j\left((ApX)_{j,i}^{w_0}\right) \times \times \sum_{j=1}^{M_0}\left((ApX)_{j,i}^{w_0}\right) \Big/ i = \overline{1,N},\ \ j = \overline{1,M_0};$$

$$SApX_i^{w_1} = \max_j\left((ApX)_{j,i}^{w_1}\right) \times \times \sum_{j=1}^{M_1}\left((ApX)_{j,i}^{w_1}\right) \Big/ i = \overline{1,N},\ j = \overline{1,M_1}. \quad (10)$$

For each class (w_1 and w_0) indicators x_j are sorted in descending order of values of $\left[SApX_i^{w_0/w_1}, \varepsilon\right]$. Thus, a formed two-tuple of indices for alternative classes: $\{X^{w_0}\}$ and $\{X^{w_1}\}$.

The job ε, the formation of tuples, and further application of formula (2), the formation of many informative features $XI_j = \left(x_j / Inf\left(x_j\right)\right) / j = \overline{1,NI}$ and calculating the information content is then the same as discussed in method 1 procedures.

In method 3 linguistic variables take values "true" ("1") or false (0). With a certain accuracy (diagnostic performance in medical applications), the approximant of the response is represented by the formula (11) (indices and variables have counterparts in (2)).

$$YB(ZB) = \bigcup_{L}^{k=1}\left(\bigcap_{N}^{i=1}(zb)\right), \quad (11)$$

where $f_i(zb) \in \{zb.\overline{z}, 1\}$; $zb \in \{ZB\}$ – logic exception.

For possible applications of the approaches described in method 1 and method 2 from (11) to proceed to the analogue of the polynomial Gabor YB^* (ZB^*) for Boolean functions in the

form of formula (12), based on analogues of arithmetic operations logical functions.

$$YB^*\left(ZB^*\right)=1-\prod_{i=1}^{N}\left(1-zb_i^*\right)^{p_k},$$

$$p_k=\{0,1\},\ zb_i^*=\{0,1\}. \qquad (12)$$

Then apply formula (5) to(10) and conclusions from the consequences.

Method 4 proposes to implement the ordering of attributes (linguistic variables) with the subsequent calculation of grades, the inclusion in an informative tuple and the calculation of informativeness similar to the previously discussed methods on the basis of hyperobject H (and/or index density PD), considered in [7], conducting exploratory clustering procedure by calculating the value of changes in the quality of clustering dH_{x_j} as the exception from consideration of the analyzed characteristic by the formula

$$dH_{x_j}=\frac{\sqrt{\left[\left(\det\left(R_{w_0,\{X\}}\right)\right)^2+\left(\det\left(R_{w_1,\{X\}}\right)\right)^2\right]}-\sqrt{\left[\left(\det\left(R_{w_0,\{X\}-x_j}\right)\right)^2+\left(\det\left(R_{w_1,\{X-x_j\}}\right)\right)^2\right]}}{\sqrt{\left[\left(\det\left(R_{w_0,\{X\}}\right)\right)^2+\left(\det\left(R_{w_1,\{X\}}\right)\right)^2\right]}}, \qquad (13)$$

where $R_{w_0,\{X\}}$; $R_{w_1,\{X\}}$ – is a covariance matrix into the corresponding classes w_0 and w_1 in the initial set $\{X\}$; $R_{w_0,\{X\}-x_j}$; $R_{w_1,\{X-x_j\}}$ – is the correlation matrix of the classes w_0 and w_1 the set $\{\{X\}-x_j\}$ (excluded sign x_j; $det(\ )$ – compute the determinant of the matrix.

Under covariance matrices here are the matrices calculated by the formulas

$$R_{w_0,\{X\}}=\frac{1}{N_0}\cdot\sum_{i=1}^{N_0}\left(x_{w_0,i}-v_{w_0}\right)\cdot\left(x_{w_0,i}-v_{w_0}\right)^T;$$

$$R_{w_1,\{X\}}=\frac{1}{N_1}\cdot\sum_{i=1}^{N_1}\left(x_{w_1,i}-v_{w_1}\right)\left(x_{w_1,i}-v_{w_1}\right)^T, \qquad (14)$$

where N_0, N_1 – is the number of objects in classes w_0 and w_1, respectively; $x_{w_0,i}$, $x_{w_1,i}$ – coordinate vector of the i-th object in the respective clusters; v_{w_0}, v_{w_1} – vectors of coordinates of the centers of the classes w_0 and w_1.

Note that dH_{x_j} can take both positive and negative values – the latter option means that after breeding the quality of the classification according to the General hyperonym H deteriorated.

The disadvantage of this method is the analysis of exception characteristic as a single representative, rather than together with some other tuples. Procedure complete enumeration of different variants of demand in this case, large computational resources are usually, with negligible loss of diagnostic quality (or lack thereof) in the end.

In conclusion, we note that:

1. In the proposed methods, the information content characteristic is determined for each "branching" of the tree of decision-making about the object or process alternative classes. Thus, from the paradigm definition, equal informative tuples linguistic variables for the full set of alternative classes (and, subsequently, the synthesis of diagnostic rules), it is proposed to move to the paradigm of determining the informational content of the basis for each hierarchy, differential division.

2. If in the formula (2) to move from $zb_i^*=\{0,1\}$ go to $zb_i^*=]0,1]$, then the binary characteristic value, go to the interval estimates of the characteristic values of membership functions in fuzzy sets or functions of belief in the theory of decision-making.

Thus, in the course of the study developed a new nonparametric methods of formation of informative tuples describing observable and/or controllable signs (linguistic variables) of the biological object (recorded, calculated, and latent, in numeric and logical metrics), which allows in conditions of semi-structured imprecise data necessary for the synthesis of diagnostic decision rules knowledge bases decision support systems in various segments of the automation of intellectual activities of decision makers on the basis of modern computer and information technologies.

References

1. Artemenko M.V., Babkov A.S. Classification of methods of forecasting the behavior of systems // Modern problems of science and education. – 2013. – № 6; URL: http://www.science-education.ru/ru/article/view?id=11527 (date accessed: 8.06.2016).

2. Artemenko M.V., Dobrovolsky I.I., Mishustin V.N. Information-analytical support of the automated classification on the basis of direct and inverse decision rules on the example of prediction of thromboembolic disease // Modern high technologies. – 2015. – № 12–2. – P. 199–205.

3. Artemenko M.V., Korenevsky N.A. Jelinkova L.A. Diagnostics of the health of the newborn through systemic analysis of pregnant indicators // Bulletin of new medical technologies. – 2003. – T. 10. – № 3. – P. 50–52.

4. Artemenko N.M. Recognition of the state of human lungs in that they produce acoustic noise // proceedings of southwest state University // Series: Management, computer engineering, computer science. Medical devices. – 2015. – № 2 (15). – P. 94–98.

5. Barsky A. B. Logical neural networks. – M.: NOU "Intuit", 2016. – 492p.

6. Vorontsov I. M., Shapovalov V.V., Sherstuk Y.M. Health. Experience in the development and justification of the application of automated systems for monitoring and srinilaya diagnosis of health disorders. – SPb.: OOO "IPK "Costa" B, 2006. – 432 p.

7. Demidova L.A., Kirakovskii V.V., Pylkin A.N. Decision-making in conditions of uncertainty. – 2nd ed. revised – M.: Hot line – Telecom, 2015. – 283 p.

8. Zagoruiko N.G., Kutnenko O.A. The division Algorithm for selecting informative subspaces of signs Institute of Mathematics SB RAS (access point http://pandia.ru/text/78/248/79351.php).

9. Zhvalevsky A.V. The Selection of informative features: setting objectives and methods of its solution // Proceedings of SPIIRAS. – 2007. – Vol. 4. – P. 416–426.

10. Lbov G.S., Startseva N.G. Logical decision functions and the question of statistical stability of solutions. – Novosibirsk: publishing house of Institute of mathematics, 1999. – 212 p.

11. Orlov A.A. the Principles of the architecture of a software platform for implementing the algorithms of the group method of accounting arguments // Control systems and machines. – 2013. – № 2. – P. 65–71.

12. The multiplicative approximation method of group accounting of arguments // The Certificate of official registration program for computer № 2007611654 from 25.04.2007

13. Research library natural science selected publications. – URL: http://sernam.ru.

14. Saaty Thomas L. Decision making with dependence and feedbacks: analytical networks. Per. s angl / Scientific. edited by A.V. Andreychikov, O.N. Andreichikova. Ed. 4. – M.: LENAND, 2015. – 360 p.

15. Handbook on mathematics for researchers and engineers // King., Corn. – M.: Science 2007, – 789 p.

16. Artemenko M.V., Dobrovolsky I.I. Automated test system training of doctors trauma-based support module expert diagnostic // 3rd international conference "Research, Innovation and education" at the SCIERO. – London, January 25–30, 2016 – P. 226–238.