ORIGINAL PAPERS

Adv Clin Exp Med 2011, **20**, 2, 205–209
ISSN 1230-025X

Henryk Kordecki[1], Mikołaj Karmowski[2], Maria Knapik-Kordecka[3], Andrzej Karmowski[4], Bohdan Gworys[5]

Method of Component Importance Evaluation in Complex Data Structure Analysis

Metoda określenia ważności parametrów w analizie danych o złożonej strukturze

[1] Institute of Computer Technology, Automatics and Robotics, Wroclaw University of Technology, Poland
[2] Department of Gynaecology and Obstetrics, Wroclaw Medical University, Wrocław, Poland
[3] Department of Angiology, Hypertension and Diabetology, Wroclaw Medical University, Wrocław, Poland
[4] 1st Department and Clinic of Gynaecology and Obstetrics, Wroclaw Medical University, Wrocław, Poland
[5] Department of Normal Anatomy, Wroclaw Medical University, Wrocław, Poland

Abstract

Background. In medical practice very often it is necessary to determine which parameter (feature, cure, environment element) is most important for diagnosis or decision making. This is the problem of complex, multi-component data analysis. The paper introduces the method that can be helpful to solve that kind of problems.
Objectives. The aim of the paper was to present the method of data components ordering according to their importance, in the process of complex medical problem analysis.
Material and Methods. The idea implemented in this paper is based on the entropy conception, known in the area of information theory. As the example Eurofit motor test results of 637 males and 425 females were analyzed. Test results contain 9 parameters for each individual, and the proposed method was used for ordering them to simplify the individual efficiency evaluation.
Results. The applied method of test components ordering gave good results. The quality of results was evaluated by comparison the ordering for males and females using correlation analysis.
Conclusions. Results obtained using the proposed method allows to state that the method can be used in cases of the multi-parameter medical data analysis when the parameter ranking is necessary (**Adv Clin Exp Med 2011, 20, 2, 205–209**).

Key words: entropy, data component ordering, motor tests.

Streszczenie

Wprowadzenie. W praktyce medycznej często zachodzi konieczność określenia, który parametr (cecha, lekarstwo, element środowiska) jest najważniejszy do określenia diagnozy lub w procesie podejmowania decyzji. Jest to problem złożonej, wieloparametrycznej analizy danych. W pracy opisano metodę, która może być zastosowana do rozwiązywania takich problemów.
Cel pracy. Prezentacja metody uporządkowania komponentów danych w zależności od ich ważności w procesie analizy złożonego problemu medycznego.
Materiał i metody. Metoda opisana w pracy opiera się na pojęciu entropii znanym teorii informacji. Jako przykład jej zastosowania wybrano wyniki testu sprawnościowego Eurofit wykonanego u 637 mężczyzn i 425 kobiet. Zawierają one 9 parametrów dla każdej osoby, a opisana metoda była wykorzystana do ich uporządkowania w celu ułatwienia oceny sprawności poszczególnych osób.
Wyniki. Przedstawiona metoda uporządkowania parametrów dała dobre wyniki. Jakość rezultatów oceniono przez porównanie uporządkowania dla kobiet i mężczyzn, wykorzystując analizę korelacji.
Wnioski. Rezultaty otrzymane z wykorzystaniem opisanej metody pozwalają sądzić, że może być wykorzystana do analizy wieloparametrycznych danych medycznych, w przypadku gdy uporządkowanie parametrów jest konieczne (**Adv Clin Exp Med 2011, 20, 2, 205–209**).

Słowa kluczowe: entropia, porządkowanie komponentów danych, testy motoryczne.

The aim of the paper is to present the method of components ordering according to their importance and level of influence on the process of complex medical problem analysis. In medical practice very often it is necessary to determine which parameter (feature, cure, environment element) is most important for decision making. The method proposed, named "empirical entropy method" should be helpful to make a right decision during the parameter importance evaluation in complex data structures. The idea implemented in this paper is based on the entropy conception, presented by Shannon [1] in the area of the information theory. The entropy conception was used untypical way in the process of company flexibility evaluation. In this paper besides the entropy conception the correlation analysis was applied [2]. The usefulness of the proposed method was tested in the process of the motor abilities analysis by EUROFIT test application [3] in the group of university students. The EUROFIT test consists of nine efficiency trials. The question is which one of them is the most important for the individual motor ability evaluation and which one is unessential. Similar approach was presented in the area of multiple attribute decision making (MADM) [4]. This process refers to making decisions (evaluation, selection) over available alternatives described by multiple attributes. In this paper the usefulness of the proposed method was evaluated by using the assumption that ordering of the test components (parameters) should be similar for males and females. So, it was assumed that the more similar the importance ordering of test components for males and females the more useful the proposed method.

Material and Methods

Test results of 637 males and 425 females were analyzed. Basic statistical parameters of EUROFIT test for both groups after modification are given in Tables 1 and 2.

Modifications of the test parameters were made according to the rule "the more the better". For the parameters 2, 7, 8, and 9 the inverses were taken into account. It is obvious that the lower time of running (in seconds) the better result. After modifications all the parameters were standardized. The importance ordering of the test parameters was made according to the principle that the more global information contains the parameter the more its importance in the process of the individual motor ability evaluation. As the measure of the amount of information included in the particular test parameter the empirical entropy was used [1, 2].

Results

The values of test parameters were treated as samples of continue random variables. For females and males, the frequency histograms (empirical probability density distributions) were built. For the use of these investigations the number of histogram classes, in each case, was established to 10. Then empirical probabilities of given parameter achieving value in particular class of the corresponding histogram were calculated according to the formula:

$$p_{ik} = \frac{l_{ik}}{m} \tag{1}$$

Table 1. Basic statistical parameters of the male group

Tabela 1. Podstawowe parametry statystyczne w grupie męskiej

Parameter	Sample size	Average	Minimum	Maximum	Standard deviation
1. Flamingo balance	637	7.52	1.00	26.00	4.54
2. Plate tapping	637	10.16	7.02	20.68	1.38
3. Sit-and-reach [cm]	637	26.97	7.00	50.00	7.92
4. Standing broad jump [cm]	637	224.89	110.00	290.00	22.88
5. Handgrip strength [kg]	637	46.58	16.00	80.00	9.66
6. Sit-ups [number/30 sec]	637	27.57	15.00	44.00	4.45
7. Flexed arm hang [sec]	637	28.34	1.05	76.01	13.81
8. Shuttle-run 10 × 5 m. [sec]	637	19.42	15.29	29.04	1.73
9. Buree's test	637	7.38	4.00	14.00	2.00

Table 2. Basic statistical parameters of the female group

Tabela 2. Podstawowe parametry statystyczne w grupie żeńskiej

Parameter	Sample size	Average	Minimum	Maximum	Standard deviation
1. Flamingo balance	425	6.71	1.00	25.00	4.03
2. Plate tapping	425	11.28	8.09	19.39	1.36
3. Sit-and-reach [cm]	425	27.90	7.00	47.00	6.58
4. Standing broad jump [cm]	425	173.06	120.00	223.00	18.31
5. Handgrip strength [kg]	425	25.02	10.00	49.00	6.17
6. Sit-ups [number/30 sec]	425	23.28	13.00	39.00	3.64
7. Flexed arm hang [sec]	425	10.84	0.53	38.56	7.87
8. Shuttle-run 10 × 5 m. [sec]	425	21.44	16.88	25.97	1.57
9. Buree's test	425	4.76	2.00	12.00	2.02

where:
m – number at individuals (m = 637 for males, m = 425 for females),
i – test parameter (component) index (i = 1…9),
k – histogram class index (k = 1…10),
l_{ik} – number of samples of i*th* test component observed in *k*th class of histogram,
p_{ik} – empirical probability of i*th* test component achieving value in k*th* histogram class. Thus the i*th* parameter value will occur in the k*th* class $p_{ik} \times m$ times.

It is also obvious that:

$$\sum_{k=1}^{10} p_{ik} = 1 \text{ , i = 1…9} \tag{2}$$

To present the idea, which is the basis of the proposed method implementation, it is necessary to explain more precisely the concept of the test component importance. According to the author opinions, the most important test component is that one for which achieving the good result is either very easy or very difficult. If someone achieved a good result in a very difficult test component it means that, he or she, is more efficient (trained) than others. The achievement of the bad result in easy test component means that given individual is less efficient than others.

For some test components, probabilities of achieving result in each histogram class are equal, for some not. If they are equal or near equal means that the empirical probability distribution of the corresponding parameter is nearly uniform and this particular parameter cannot be used for differentiation of the motor ability. The more, the empirical probability distribution of the particular parameter is different from uniform the more important is that parameter in the process of the individual motor ability comparisons. This fact is used in proposed method for the importance evaluation of test components.

Separately for males and females, the amount of the information contained in i*th* data sample related to i*th* test parameter was calculated according to the formula below [5]:

$$I_i = \sum_{j=1}^{k} m \times p_{ij} \times \log_2 (\frac{1}{p_{ij}}) \tag{3}$$

Value H_i: $$H_i = \frac{I_i}{m} = \sum_{j=1}^{k} p_{ij} \times \log_2 (\frac{1}{p_{ij}}) \tag{4}$$

is the average amount of information per one element of the i*th* test parameter and can be called the empirical entropy of that parameter (component) [5]. The entropy H_i ($H_i \geq 0$), achieves maximum when probabilities p_{ij} are equal for each "j", what means that the probability distribution of the i*th* component is uniform [6]. It denotes that the test component with the maximal entropy corresponds to that kind of situation in which probabilities of achieving god and bad results are the same for efficient and inefficient individuals. According to author opinions, that kind of component is less important for the efficiency evaluation then that one with smaller entropy. Thus, the less entropy has the test parameter the more its importance in the motor abilities evaluation process. All test components were ordered according to the increasing entropy H_i, separately for males and females. Results are shown in Table 3.

Table 3. Eurofit test components importance evaluation for males and females

Tabela 3. Ocena ważności składników testu Eurofit dla mężczyzn i dla kobiet

	ENTRM	RANKM	ASYMM	ENTRF	RANKF	ASYMF
1. Flamingo balance	3.858	2	0.843	3.825	3	1.174
2. Plate tapping	4.289	8	–0.387	4.258	8	0.087
3. Sit-and-reach [cm]	4.197	4	–0.005	4.216	7	–0.203
4. Standing broad jump [cm]	4.237	5	–0.576	4.199	6	–0.053
5. Handgrip strength [kg]	4.241	6	0.131	4.161	5	0.404
6. Sit-ups [number/30 sec]	4.145	3	0.272	2.976	2	0.297
7. Flexed arm hang [sec]	4.328	7	0.407	4.054	4	1.156
8. Shuttle-run 10 × 5 m. [sec]	4.269	9	–0.462	4.267	9	0.173
9. Buree's test	2.973	1	0.848	2.822	1	1.006

Discussion

In Table 3 results of the test components importance evaluation are displayed. Columns ENTRM and ENTRF contain entropy values and columns RANKM and RANKF contain ranks obtained according to the increasing entropy for males and females respectively.

The small entropy value is characteristic for difficult as well as for easy test parameters. In order to verify the difficulty level, skew coefficients "A" were calculated for each parameter according to the formula 5 [7]. The results are shown in the columns ASYMM and ASYMF of Table 3 for males and females respectively.

$$A = \frac{\overline{X_3}}{\sigma^3} \tag{5}$$

where:

$\overline{X_3}$ – third range moment calculated for data sample of the each test component,

σ – standard deviation of each sample.

Bigger absolute value of the skewness coefficient denotes bigger distribution asymmetry. If A > 0 the distribution is said to be positively skewed ore skewed to the right. In this case the probability of achievement the good result is small and the trial of the test can be considered to be difficult. The good result shows good motor efficiency of the individual. If the value of A is negative the distribution is called to be negatively skewed and the trial of the test is considered to be easy. Examples of histograms for A > 0 and for A < 0 are shown in Fig. 1 and Fig. 2 respectively.

Following the idea that the more similar the importance ranking of the test components for males

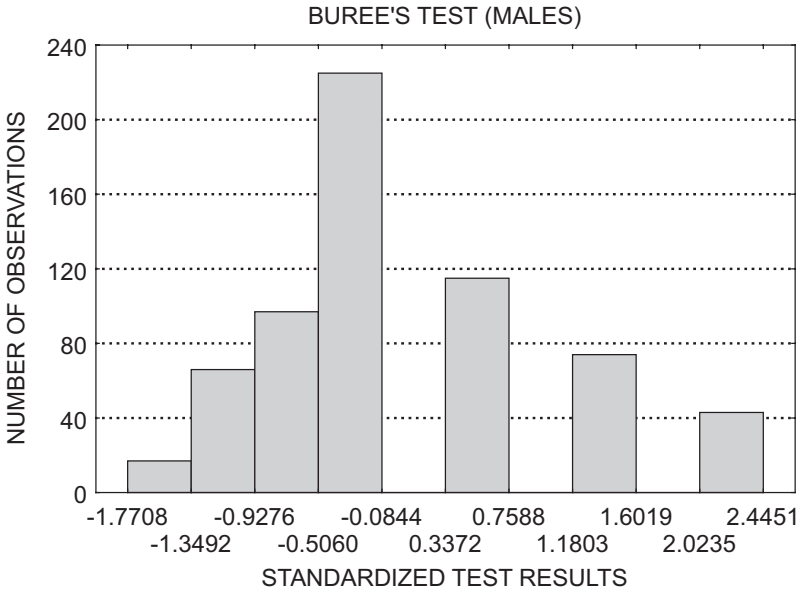

Fig. 1. Histogram for Buree's test skewed to the right (A > 0)

Ryc. 1. Histogram dla składnika testu „bieg wytrzymałościowy" z prawostronną asymetrią

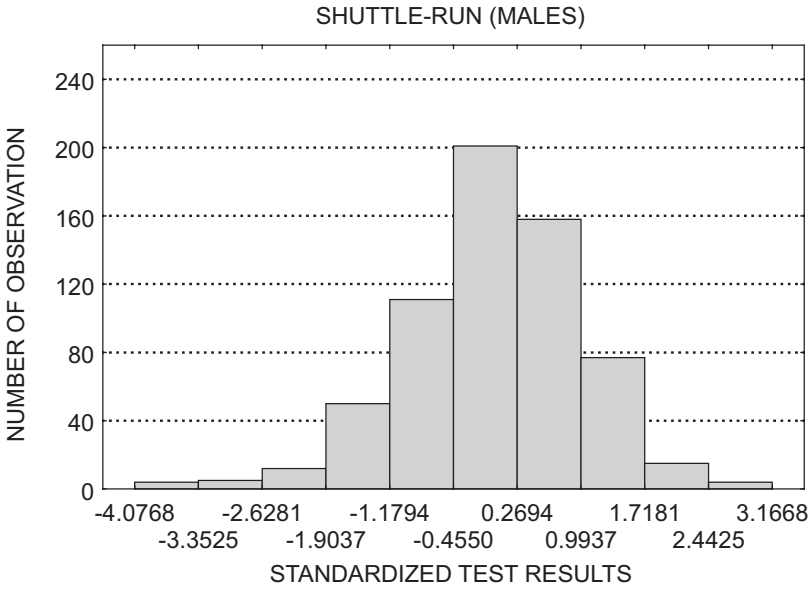

Fig. 2. Histogram for shuttle-run skewed to the left (A < 0)

Ryc. 2. Histogram dla składnika testu „bieg wahadłowy" z lewostronną asymetrią

and females the more useful the proposed method, Spearman's correlation coefficient "R" between ranks, shown in Table 3 (RANKM and RANKF columns), was calculated. The value of R obtained, was very high, R = 0.817. This result confirmed the usefulness of the presented method.

The authors concluded that the original method of test components ordering, according to their importance for motor efficiency evaluation, called "empirical entropy" method gave good results. The ranking of test components obtained, using the presented method, was very similar for males and females. This result confirmed the method usefulness, if the assumption, that the good method should give the same results independently on the sex, will be accepted.

Results obtained using the "empirical entropy" method allows to say that the method can be used in all cases of the multi-component medical data analysis, when the criterion of component ordering, will be the amount of information contained in the each component.

References

[1] **Shannon CE:** A mathematical theory of communication. Bell System Tech J 1948, 27, 379–423, 623–656.
[2] **Harnet DL:** Statistical methods. Addison-Wesley Publishing Company, Inc. Philippines 1982, 718–728.
[3] Council of Europe. Committee for the development of sport. Eurofit, European Test of Physical Fitness 1988.
[4] **Hosseinzadeh L, Fallahnejad R:** Imprecise Shannon's entropy and multi attribute decision. Entropy – Open Access J 2010, 12, 53–62.
[5] **Shanmugam K Sam:** Digital and analog communication systems, Jon Wiley & Sons Inc. 1979, 138–156.
[6] **Harremoes P:** Maximum entropy on compact groups. Entropy – Open Access J 2009, II, 222–237.
[7] **Azzalini A, Capitanio A:** A statistical applications of the multivariate skew normal distribution. J R Stat Soc 1999, 81 (3), 579–602.

Address for correspondence:

Henryk Kordecki
Institute of Computer Technology, Automatics and Robotics
Wroclaw University of Technology
27 Wybrzeże Wyspiańskiego St.
50-370 Wrocław
Poland
Tel.: +48 71 320 29 61, +48 665 224 555
E-mail: henryk.kordecki@pwr.wroc.pl

Conflict of interest: None declared

Received: 10.02.2011
Revised: 31.03.2011
Accepted: 1.04.2011