COMPARISON OF WAYS OF NORMALIZATION AT CLASSIFICATION OF INITIAL DATA

Farzaliyev Y.Z.

Azerbaijan Scientific-Research and Design-Prospecting Institute of Energetic
AZ1012, Ave. H.Zardabi-94,

e-mail:yuszey2002@yahoo.com

ABSTRACT

Normalization at classification of initial data is an indispensable condition of an opportunity of comparison of technical and economic characteristics of the equipment and devices of power supply systems. In work, results of comparison of efficiency of ways of normalization of quantitative estimations of analyzed characteristics used in practice are resulted.

Keywords. Normalization, classification, the equipment, characteristics, the importance attributes.

I. INSTRUCTION

The automated classification of the equipment, its nameplate data and data on conditions of operation, refusals, test and repair on the set versions of attributes (VA) is widely used at the decision of some exploitation problems, volume number of problems about expediency of carrying out of maintenance service and repair. For example, it is necessary to establish most both the least reliable and economic power units. Value of the first is necessary at the decision of problems on distribution of loading, switching-off in a reserve, emergency switching-off at system failures and so forth. For the least reliable and economic power units carrying out of corresponding maintenance service and restoration of deterioration is required.

Thus, problems about the importance of attributes solved, as a rule, at an intuitive level, and ranging VA on the importance - is subjective, that is why it is often erroneous. Difficulties of an estimation of importance VA is caused, first, by a natural variety of scales and scale of measurement of attributes, their interrelation and a various orientation of influence.

According to the accepted division of a scale happen quantitative (for continuous random variables, for example, residual service life, resistance to a direct current, loading, extent of transmission lines, number of short circuits, etc.), interval (for example, month of year, rated voltage and capacity, etc.), nominal, establishing quality standards (for example, quality of repair, type and conditions of the equipment, etc.).

Methodology of ranking of importance VA with interval and nominal scales of measurement has been considered by us in [1]. If for interval and nominal scales number VA fixed, at a quantitative scale of measurement classification demands development of special approaches. In practice, the quantitative scale, as a rule, transformed in interval. The number of intervals (i.e. VA) established subjectively, proceeding from features of a solved problem. Character of a divergence of average values of realizations VA among themselves and all data set does not stipulate, though divergences can have casual character. At classification of data it leads to loss of the information and finally to growth of risk of the erroneous decision. For prevention of these consequences, it is necessary to develop a method of classification of data, which would consider expediency of classification and overcame noted above difficulty.

II. TRANSFORMATION ALGORITHM OF INITIAL DATA.

Provides following sequence of calculations:

1. Maintenance of a uniform orientation of influence. As «an orientation of influence», understand character of change of properties of the equipment (for example, reliability and profitability of power units) at increase or reduction of their quantitative characteristics (we shall agree to name their parameters). For example, the increase in the specific charge of conditional fuel (C_F) testifies to decrease in profitability of work owing to deterioration of a technical condition of the capital equipment of power station. The same can be approved at increase in the charge of the electric power at own needs (P_{ON}). But the more operating ratio of the established capacity (R_C), the reliability and profitability of work of the power unit above. In other words, C_F and R_C have a various orientation. The automated comparison of such parameters leads to erroneous decisions. Recognition of an orientation of influence at a small number of parameters is often accessible manually. With increase in number of parameters, the probability of a mistake increases. In the automated mode, recognition of distinction of an orientation of influence offered to spend by construction of a correlation matrix. As a whole, this matrix will appear interrelation of parameters necessary at estimation. For maintenance of a uniform orientation of parameters, it is enough to choose as the first (control) parameter of a matrix a parameter with known character of change at change of properties of object. Then the first line (column) of a matrix on a negative sign on factors of correlation will allow establishing number and conditional number of the parameters having other orientation, than a control parameter.

2. Classification of data to the interconnected attributes is inexpedient, since also leads to loss of the information and growth of risk of the erroneous decision. Association of the interconnected versions of attributes allows to lower essentially their number and to simplify calculations. The quantitative characteristic of interrelation of attributes can be calculated in the form of estimations of factors of correlation $(K^*_{i,j})$. Than number of classified objects (m_Σ) It is less, that casual character $K^*_{i,j}$, where i≠j; i=1, m_Σ; j=1, m_Σ, It is shown in a greater degree, both on size, and on a sign. In this connection dependence between attributes with probability not smaller than (1-α), where α - the significance value, can be confirmed by the control of performance of following inequalities: $K^*_{i,j} < \underline{K}_{i,j,\alpha}$ at $K^*_{i,j} < 0$ and $K_{i,j} > \overline{K}^*_{i,j,\alpha}$ at $K_{i,j} > 0$, where $\underline{K}_{i,j}$ and $\overline{K}_{i,j}$ - critical values of factor of correlation at $F^*(K_{i,\,j})$, according to equal α and (1-α), where $F^*(K_{i,j})$ – statistical function of distribution $K_{i,j}$. The estimation $\underline{K}_{i,j}$ also $\overline{K}_{i,j}$ is spent in following sequence:

2.1. Two samples from m_Σ are modeled random variables ξ with uniform distribution in an interval [0,1];

2.2. Count factor of correlation between realizations samples;

2.3. Items 1 and 2 repeat N time;

2.4. On N to realizations $K^*_{1,2}$ is under construction $F^*(K_{1,2})$;

2.5. Are defined $\underline{K}_{1,2}$ and $\overline{K}_{1,2}$ for set α.

If to assume, that difficulties with various orientations VA eliminated, i.e. all realizations $K^*_{i,j}$ will be positive, the size $\overline{K}_{i,j}$ is necessary for practical calculations only. Some results of calculations $\overline{K}_{i,j}$ for of some α and m are resulted in table 1, and in figure 1 experimental and theoretical dependences $\overline{K}_{i,j} = f(m)$ for most often used value are resulted α=0,05.

Table 1

Critical values of factors of correlation independent samples

n	Significance value			
	0,95	0,975	0,99	0,995
	$\overline{K}$	$\overline{K}$	$\overline{K}$	$\overline{K}$
3	0,989	0,998	0,999	0,999
5	0,778	0,864	0,937	0,959
8	0,632	0,712	0,726	0,847
10	0,514	0,608	0,689	0,730
20	0,393	0,452	0,506	0,534
30	0,301	0,357	0,424	0,446
50	0,220	0,274	0,320	0,377

It is established, that for $m_\Sigma \geq 3$ size $\overline{K}_{1,2}$ with reliability not less than 0,99 can be calculated under the formula $\overline{K}_{1,2} = 1.79/\sqrt{m_\Sigma}$. For example, at m_Σ=20 sizes $\overline{K}_{1,2} = 1.79/\sqrt{20} = 0.40$, and a divergence with result of modeling do not exceed 2,5 %

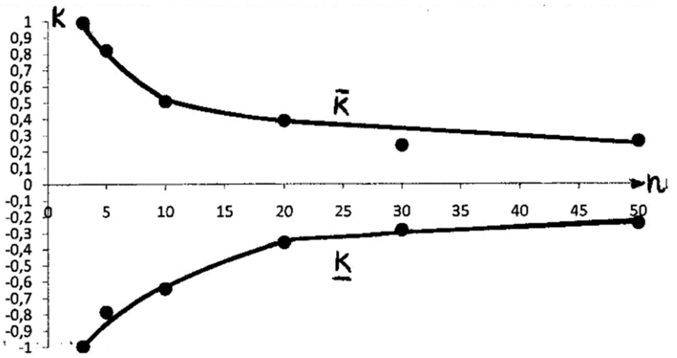

Fig.1. Laws of change of critical value $\overline{K}_{1,2}$ in function m_Σ For α=0,05.

3. Overcoming of influence of various units of measure and scale of parameters is reached by normalization (standardization) of quantitative estimations

Normalization spent on one of following formulas [2]:

$$Z = \frac{X}{\overline{X}};\ \frac{X}{X_{max}};\ \frac{X-\overline{X}}{\sigma};\ \frac{X-\overline{X}}{L}$$

where $\overline{X} = m^{-1}\sum_{i=1}^{m} X_i$; X_{max}=max {X_1; X_2;..., X_m}; X_{min}=min {X_1; X_2;..., X_m}, $\sigma = \sqrt{\frac{(X-\overline{X})^2}{m-1}}$; L=X_{max}-X_{min};.

The Importance of these formulas does not stipulate.

As it has noted been above, normalization of initial data is necessary for comparison of attributes. We shall specify, about what comparison there is a speech. For this purpose, it is enough to consider only the first stage of recommended algorithm of classification of initial data [3]. The essence of the first a stage consists in the following. Let given m_Σ objects (for example, power units or transformers, or switches) and their basic characteristics. It is required to range these objects by way of increase of reliability and profitability of their work, for what:

- Realizations of each of quantitative characteristics we shall consider as set of random variables;
- We calculate a number of their statistical parameters. Namely: average arithmetic value $M^*_\Sigma(\Pi_i)$, minimal Пi,min and maximal Пi,max values, disorder $L^*_\Sigma(\Pi_i)$ and an average quadratic deviation $\sigma^*_\Sigma(\Pi_i)$ under formulas:

$$M^*_\Sigma(\Pi_i) = m_\Sigma^{-1}\sum_{j=1}^{m_\Sigma}\Pi_{i,j}\,; \qquad \Pi_{i.\min} = \min\{\Pi_i\}_{m_\Sigma}\,; \qquad \Pi_{i,\max} = \max\{\Pi_i)_{m_\Sigma}\,; \qquad L^*_\Sigma(\Pi_i) = (\Pi_{i,\max} - \Pi_{i,\min})\,;$$

$$\sigma^*_\Sigma(\Pi_i) = \sqrt{\frac{[\Pi_i - M^*_\Sigma(\Pi_i)]^2}{(m_\Sigma - 1)}}$$

- Realizations for which $\Pi_i > M^*_\Sigma(\Pi_i)$ carry to the first sample (to the first version i- th an attribute). Realizations, for which $\Pi_i < M^*_\Sigma(\Pi_i)$ with i=1,nΣ – to the second sample (accordingly to the second version i- th an attribute). Such classification of data is widely used in practice, physically proved;
- For both samples (v) each set average arithmetic values $M^*_{v,1}(\Pi_i)$ and $M^*_{v,2}(\Pi_i)$ with i=1,nΣ are calculated. We shall notice, that essential distinction $M^*_{v,1}(\Pi_i)$ and $M^*_{v,2}(\Pi_i)$ is caused by distinction of number of both elements samples ($m_{v,1,i} \neq m_{v,2,i}$);
- For everyone (i=1,nΣ) data sets we define sample, for which divergence $\Delta M^*_v(\Pi_i)$ the greatest, i.e. $\Delta M^*_v(\Pi_i)$=max$\{\Delta M^*_{v,1}(\Pi_i)\,;\Delta M^*_{v,2}(\Pi_i)\}$, where $\Delta M^*_{v,1}(\Pi_i) = [M^*_{v,1}(\Pi_i) - M^*_\Sigma(\Pi_i)]$, and $\Delta M^*_{v,2}(\Pi_i) = [M^*_\Sigma(\Pi_i) - M^*_{v,2}(\Pi_i)]$;
- Define the greatest value $\Delta M^*_v(\Pi)$ among nΣ values $\Delta M^*_v(\Pi_i)$. Here we collide with distinction of dimension and scale of attributes.

In the subsequent, we shall consider efficiency of following transformations

$$\delta\Pi_{i,1} = \frac{\Pi_i}{M^*_\Sigma(\Pi_i)}\,;\; \delta\Pi_{i,2} = \frac{\Pi_i}{L^*_\Sigma(\Pi_i)}\,;\; \delta\Pi_{i,3} = \frac{\Pi_i - M^*_\Sigma(\Pi_i)}{M^*_\Sigma(\Pi_i)}\,;\; \delta\Pi_{i,4} = \frac{\Pi_i - M^*_\Sigma(\Pi_i)}{L^*_\Sigma(\Pi_i)}$$

III. RESULTS OF COMPARISON WAYS OF NORMALIZATION INITIAL DATA

If to consider the above-stated it is easy to conclude, that without normalization, ranging of objects on the importance at number of attributes nΣ>1 it is labour consuming and with growth nΣ Labour input increases. In the illustrative purposes in table 2 two characteristics of power, units are resulted and it is required to range these power units on the importance.

Table 2

Monthly average data on work PU

Parameter	Conditional number PU							
	1	2	3	4	5	6	7	8
Specific charge of conditional fuel	374,6	371	368,4	369,7	336,7	373,9	363	374,2
Charge of the electric power on own needs	4,1	4,4	4,0	3,9	3,5	4	3,7	3,5

For normalization we shall define, corresponding statistical parameters of the sets resulted in table 2. Results of calculations given in table 3

Table 3

Results of calculations of sample parameters

Parameter	$M^*_\Sigma(\Pi_i)$	$\Pi_{i,\min}$	$\Pi_{i,\max}$	$L^*_\Sigma(\Pi_i)$
Specific charge of conditional fuel	366,4	336,7	374,6	37,9
Charge of the electric power on own needs	3,89	3,5	4,4	0,9

In tables 4 and 5 results of normalization, accordingly, estimations of the specific charge of conditional fuel (C^*_F) and the charge of the electric power on own needs P^*_{ON}) are resulted.

Table 4

Results of transformation of estimations of the specific charge of conditional fuel

Conditional PU	Type of transformation			
	$\delta C_{F,1}$	$\delta C_{F,2}$	$\delta C_{F,3}$	$\delta C_{F,4}$
1	1,021	9,88	0,021	0,21
2	1,012	9,79	0,012	0,12
3	1,005	9,72	0,005	0,05
4	1,009	9,75	0,009	0,09
5	0,919	8,88	-0,081	-0,79
6	1,020	9,87	0,020	0,20
7	0,991	9,58	-0,009	-0,09
8	1,021	9,87	0,021	0,21

Table 5

Results of transformation of estimations of the charge of the electric power for own needs

Conditional PU	Type of transformation			
	$\delta P_{ON\ 1}$	$\delta P_{ON.,2}$	$\delta P_{ON.,3}$	$\delta P_{ON.,4}$
1	1,054	4,58	0,054	0,24
2	1,131	4,9	0,131	0,58
3	1,028	4,44	0,028	0,12
4	1,003	4,33	0,003	0,01
5	0,900	3,89	-0,10	-0,43
6	1,028	4,44	0,028	0,12
7	0,951	4,11	-0,049	-0,21
8	1,900	3,89	-0,10	-0,43

According to the sequence of classification of initial data stated above, we shall divide samples with identical transformation of realizations of random variables on two groups:

- for the first transformation the first group includes realizations $\delta\Pi_{i,1}>1$, and the second group – realizations $\delta\Pi_{i,1}<1$
- for the second transformation the first group includes realizations $\delta\Pi_{i,2} > \Pi_i / L^*_\Sigma(\Pi_i)$, and the second group – realizations $\delta\Pi_{i,2} < \Pi_i / L^*_\Sigma(\Pi_i)$
- for the third and fourth transformation is, accordingly, positive (+) and negative (-) values $\delta\Pi_i$.

In semantic aspect, the first group is a group of "bad" power units, and the second – from "good".

Let's define average value of realizations of each of samples for both parameters and four kinds of transformation δΠi and divergences $M^*_\Sigma(\delta\Pi_i)$ with $M^*_{v,j}(\delta\Pi_i)$, i.e. we shall define $\Delta M^*_{\Sigma,j}(\delta\Pi_i) = M^*_\Sigma(\delta\Pi_i) - M^*_{v,j}(\delta\Pi_i)$. Results of calculations are resulted in table 6.

It is obvious, that the more differs $M^*_\Sigma[\Pi_i]$ from $M^*_v[\Pi_i]$, the importance of sample above. This parity accepted to criterion of the importance of sample.

Table 6

Estimations of a deviation of average value of realizations samples from average value of population $\Delta M^*_{\Sigma,j}(\delta\Pi_i)$

Parameter	Groups	Samples of random numbers $\delta_{(\Pi i)}$			
		j=1	j=2	j=3	j=4
Specific charge of conditional fuel	1 2	0,015 0,045	0,15 0,44	0,015 0,045	0,15 0,44
Charge of the electric power on own needs	1 2	0,045 0,083	0,21 0,36	0,045 0,083	0,21 0,36

Analysis of given tables 4, 5 and 6 allows to conclude:

1. Tables 4 and 5 testify that transformations of estimations of the specific charge of conditional fuel and the charge of the electric power on own needs $\delta\Pi_{i,1}$ and $\delta\Pi_{i,2}$ though demand less calculations, but do not solve one of the main tasks of normalization – distinction of scale of measurement.
2. The greatest value of a divergence $\Delta M^*_{\Sigma,1}(\delta\Pi_i) = \Delta M^*_{\Sigma,3}(\delta\Pi_i)$ takes place for sample of realizations of the charge of the electric power for own needs of power units (0.083), and at j=2 and j=4 – for sample of realizations of the specific charge of conditional fuel of power units. Such divergence speaks distinction in reflection of statistical parameters of samples. At j=1 and j=3 average value $M^*_\Sigma(\Pi_i)$ is considered only, and at j=2 and j=4 – average value and disorder. With increase, $M^*_\Sigma(\Pi_i)$ the relative size of a deviation $\delta\Pi_i$ depends on type of a parameter. For example, for the specific charge of conditional fuel the size of a deviation is measured in terms of percent, and for average capacity – in tens percent. With increase in scope $L^*_\Sigma(\Pi_i)$ the relative size of a deviation $\Delta M^*_{\Sigma,j}(\delta\Pi_i)$ increases. The factor of correlation here is significant.
3. Hence, for recommended algorithm from the considered four variants of transformation it is expedient to use only a variant with $\delta\Pi_i = [\Pi_i - M^*_\Sigma(\Pi_i)]/L^*_\Sigma(\Pi_i)$

CONCLUSION

1. In the practice for the classification of multivariate data, are used various methods of normalization of quantitative estimations of the attributes describing object of research.
2. Among transformations used in practice by the most effective it is necessary to consider transformation, for which $\delta\Pi_i = [\Pi_i - M_\Sigma(\Pi_i)]/L_\Sigma(\Pi_i)$

REFERENCES

1. Farhadzadeh E.M., Farzaliyev Y.Z., Muradaliyev A.Z. Method and algorithm of the choice optimum number attributes describing reliability of the equipment of electro installations.

Journal: « Reliability: Theory&applications+. R&RATA (Vol.9 No.1 (32 2014, May, USA, p.73-80

2. Bureeva N.N. Multivariate the statistical analysis with use ППП “STATISTICA”. The methodical material under the program of improvement of professional skill. «Application of software in scientific researches and teaching of mathematics and mechanics». Nizhniy Novgorod, 2007, 112 p.