Algorithm of Taxonomy: Method of Design and Implementation Mechanism

N V Shalanov¹, A A Aleldinova²

¹Siberian University of the Consumer Cooperation, 26, Karla Marksa Ave., Novosibirsk, 630087, Russia
²Novosibirsk State Technical University, 20, Karla Marksa Ave., Novosibirsk, 630073, Russia

E-mail: ec_progn@sibupk.nsk.su, aletdinova@corp.nstu.ru

Abstract. The authors propose that the method of design of the algorithm of taxonomy should be based on the calculation of integral indicators for the estimation of the level of an object according to the set of initial indicators (i.e. potential). Their values will be the values of the projected lengths of the objects on the numeric axis, which will take values [0,100]. This approach will reduce the task of multidimensional classification to the task of one-dimensional classification. The algorithm for solving the task of taxonomy contains 14 stages; the example of its implementation is illustrated by the data of 46 consumer societies of the Yakut Union of Consumer Societies of Russia.

1. Introduction
The theory and practice of taxonomy was developed in the 60s of the XX century after the introduction of information and communication technologies. It is stated in the works by R.R. Sokal, G.G. Simpson, P. H. A. Sneath, N. Jardine, R. Sibson and others [1-4]. At present, the interest in taxonomy has increased, algorithms are being created and improved, new fields of application are emerging, interesting ways of achieving production and economic objectives are appearing.

The main idea of solving taxonomy tasks is as follows: there is a certain set of objects; it is necessary to divide them using a certain decision rule into a specified or unspecified number of classes. With the increase in the dimensionality of data, there are no significant changes in the algorithm of solving the task, only the volume of data processing grows.

Methods of taxonomy were preceded by methods of crossed classification. The further research showed that the principles of mathematical logic underlying the methods of crossed classification proved problematic when applied to empirical material. Empirical studies have shown the existence of natural types of phenomena that unite individual phenomena with a large number of features, for which it is impossible to find an exact definition expressed through a small set of coinciding features. In some cases, objects can be attributed to a certain type, although they lack some features or have the same features as those used in the formation of other groups. These problems led to the need to create new principles of multidimensional classification different from the classical methods.

New principles of classification, constituting the backbone of taxonomy tasks, can be described as follows: the classification of objects is not carried out sequentially by each individual feature, but simultaneously by multiple features. This set of features is given coordinates, and they form a "feature
space”. If there are n features, any object is considered as a point in the n-dimensional space of features and the classification task is reduced to identifying concentrations of these points (and, consequently, objects) in this space. The identification of these concentrations is described differently in various publications, for example, the works of A. Bouguettaya, Q. Yu, X. Liu, X. Zhou, A. Song, M. Sarstedt, E. Mooi [5, 6]. What is common to all the algorithms is that taxa (or clusters) are formed on the basis of “similarity” between objects sharing a large number of features. Single features included in the description of an object in the feature space cannot become a condition for attributing an object to a taxon. This is determined only by the whole complex of features simultaneously.

Thus, one can come to the following conclusion: most of the methods of constructing taxonomy algorithms presuppose the solution of the problem in two stages: first, the number of necessary taxa is determined; then the centers of the desired taxa are set, based on the preliminary analysis of the empirical data. These steps are quite difficult to implement if there is a large amount of data.

Besides, while setting initial number of taxa, one gets certain errors in determining their objectively existing number. With such artificial approach, there can be a risk of incorrect setting of taxa centers (even on the basis of a step-by-step iteration), which eventually leads to errors in grouping of objects into clusters.

2. Results
Let us propose the following approach in solving taxonomy problems, which overcomes the shortcomings of the classical types of algorithms.

2.1. Method of construction. Let multidimensional objects be described by features \( x_1, x_2, ..., x_n \). The initial information for \( N \) objects is described by elements \( x_{ij} \), where \( j \) is the feature number, and \( i \) is the object number. In the n-dimensional feature space, the state of any object is presented by a point whose coordinates are the values of the features describing it.

If we are talking about objectively existing taxa in the feature space, they are concentrations of points. The task of the taxonomy is to identify these concentrations. In contrast to the types of taxonomy algorithms described above, let us propose to construct an axis in this n-dimensional feature space with a beginning which coincides with the origin of the feature space and set a scale on this axis.

However, the center of an axis does not yet determine its direction. Therefore, it is necessary to take another point through which the desired axis passes. The second point can be a point which has as its coordinates the reference values of features \( X^* = (x_{11}^*, x_{21}^*, ..., x_{n1}^*) \), and an object possessing reference feature values will be called the reference one. According to the proposed approach, the reference object can be virtual, and in most cases it is, and all other objects are compared with it.

The values of the projected lengths of objects on the axis will represent the integrated indicator for the objects level estimation based on a set of initial indicators. The integrated estimation of a reference object level corresponds to 100 and the value of this indicator for any object of the population under study will be measured by the value on the scale of this axis provided that the zero level corresponds to the object having zero feature values, and 100 is related to the reference object.

The main goal of the proposed approach is to create a formula describing the axis to which the objects of the population are projected. Let

\[
y_i = \sum_{j=1}^{n} a_j z_{ij},
\]

where \( y_i \) is the value of the potential function for the \( i \)-th object, estimated by the complex of features; \( z_{ij} \) are the standardized values of the initial features calculated by formula:

\[
z_{ij} = \frac{x_{ij}}{\sigma_j}
\]

where \( \sigma_j \) - the standard deviation of sign \( x_j \); \( \alpha_j \) – the weight of the \( j \)-th feature in the potential function, calculated by the formula:
where \( z_j \) is the standardized value of the reference value of feature \( x_j \), calculated by the formula:

\[
z_j = \frac{x_j}{\sigma_j}
\]

The level of potential of the reference object is defined as:

\[
y^* = \sum_{j=1}^{n} \alpha_j z_j^*
\]

To give an integrated estimation of the level of potential of the object, it is necessary to compare its state with the reference value. The integrated estimation of potential level \( C_i \) is found from the proportion:

\[
y^* - \frac{100}{y^* - C_i}.
\]

Therefore:

\[
C_i = \frac{y_i}{y^*} \cdot 100.
\]

The formula can be presented in an expanded form:

\[
C_i = \frac{\sum_{j=1}^{n} \alpha_j x_{ij}}{\sum_{j=1}^{n} \alpha_j z_j^*} \cdot 100.
\]

or:

\[
C_i = \frac{\sum_{j=1}^{n} \alpha_j x_{ij}}{\sum_{j=1}^{n} \alpha_j \sigma_j} \cdot 100.
\]

The value of \( C_i \) corresponds to the value of the potential of the \( i \)-th object on the constructed scale. The proposed scale enables one to visualize an integrated estimation of the potential of a multidimensional object.

Let us further propose solving the taxonomy task according to one classification and the most informative criterion, i.e. the level of potential. And, thus, the problem of multidimensional classification is reduced to one-dimensional. At the same time, by substantially simplifying the solution of the problem, it is necessary to avoid the loss of information about the objects of research and achieve a high level of correctness in solving the taxonomy tasks.

Practically proposed one-dimensional classification can be carried out by ordering objects according to the level of the potential in an ascending or descending order. Since the level of potential, according to this approach, is represented by a projected length of an object on the axis in the \( n \)-dimensional feature space, after geometrical construction of this axis with the origin and scale, values of the potential levels of the objects of the population are plotted on it. Concentrations of points, which represent the potentials of objects on the axis, are objectively existing taxa. By numbering points on the axis according to the numbers of the objects of the population, objects that fall into one or another taxon are easily restored. Thus, by means of the one-dimensional classification, the taxa - classes of multidimensional objects - are identified.

Let the population of \( N \) multidimensional objects of research be described by the system of features \( x_1, x_2, \ldots, x_n \). Information on all \( N \) objects in the context of \( n \) features is reduced to the data "object-feature", whose elements \( x_{ij} \) are the values of the \( j \)-th feature of the \( i \)-th object. To solve the taxonomy task, it is necessary to carry out the following actions.
2.2. Algorithm for solving the taxonomy task.

1. Calculate mean values of features:
\[ \bar{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n} \]

2. Calculate the mean square deviations of features:
\[ \sigma_j = \sqrt{\frac{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}{n}} \]

3. Calculate the standardized values of features:
\[ z_{ij} = \frac{x_{ij}}{\sigma_j} \]

4. Determine reference values of indicators using expert valuation:
\[ X^* = (x_{1}^*, x_{2}^*, \ldots, x_{n}^*) \]

5. Calculate the standardized values of references:
\[ z_j^* = \frac{x_j^*}{\sigma_j} \]

6. Calculate the weights of the features in the potential function:
\[ \alpha_j = \frac{z_j^*}{\sqrt{\sum_{i=1}^{n} (z_j^*)^2}} \]

7. Construct the potential function:
\[ y = \sum_{j=1}^{n} \alpha_j z_j \]

8. Calculate the values of the potential function for each object in the population:
\[ y_i = \sum_{j=1}^{n} \alpha_j z_{ij} \]

9. Calculate the reference value of the potential function:
\[ y^* = \sum_{j=1}^{n} \alpha_j z_j^* \]

10. Calculate the potentials for each object:
\[ C_i = \frac{y_i}{y^*} \cdot 100 \]

11. Construct an axis with a start and a scale.

12. Apply the values of the potentials to the axis, numbering the points on the axis according to the number of the objects in the population.

13. The concentration of points reveals the number of objects which falls into taxa.

14. By this numbering, objects that fall into one or another taxon, as well as complete information about each object in the context of the initial features, are easily restored.

Thus, the proposed algorithm for solving the taxonomy problem contains 14 stages. However, the proposed algorithm requires a practical illustration. In other words, it is required to illustrate the mechanism for implementing the algorithm for solving the taxonomy problem on a specific set of multidimensional objects.

3. Discussion
Let us take 46 consumer societies of the Yakut Union of Consumer Societies as an example of a population of research objects, and there will be five features which reflect the social mission of consumer cooperation (level of social potential): \( x_1 \) – the amount of social benefits, thousand rubles (values are given in Russian Rubles (in 2017, 1 USD≈60 Rub).); \( x_2 \) – the amount of purchasing per capita, rubles; \( x_3 \) – the amount of services rendered to the population, thousand rubles; \( x_4 \) – number of newly created jobs; \( x_5 \) – number of people involved in temporary work, people.
The initial data are presented in table 1 and are located at https://ciu.nstu.ru/kaf/persons/73783/data_itbi. Calculations of variables $z_{ij}, y_i, C_i$ are shown in table 2 at https://ciu.nstu.ru/kaf/persons/73783/data_itbi. The subsequent calculations are in tables 1 and 2.

Table 1. Parameters for calculating the level of social potential of organizations

| Indicator | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ |
|-----------|-------|-------|-------|-------|-------|
| $\sigma_j$ | 27.80 | 295.34 | 88.63 | 8.64 | 41.82 |
| $x^*$ | 130.00 | 1000.00 | 62.00 | 12.00 | 22.00 |
| $z_j$ | 4.6755 | 3.3860 | 0.6996 | 1.3892 | 0.5260 |
| $\alpha_j$ | 0.7790 | 0.5642 | 0.1166 | 0.2315 | 0.0876 |
| $\alpha_j z_j^*$ | 3.6423 | 1.9102 | 0.0815 | 0.3215 | 0.0461 |

Let us find the sums:

$$\sum_{j=1}^{n} (z_j)^2 = 36.0210,$$

$$\sum_{j=1}^{n} (\alpha_j z_j^*) = 6.0018.$$

The number of the organization is determined from the initial data of Table 1 (https://ciu.nstu.ru/kaf/persons/73783/data_itbi).

Table 2. Ordered sequence of organizations by level of potential

| Level potential $C_i$ | Organization number | Level potential $C_i$ | Organization number |
|------------------------|---------------------|------------------------|---------------------|
| 12.427                 | 1                   | 59.412                 | 11                  |
| 15.360                 | 46                  | 59.838                 | 9                   |
| 18.495                 | 2                   | 60.526                 | 10                  |
| 22.118                 | 27                  | 60.543                 | 20                  |
| 28.531                 | 22                  | 61.436                 | 33                  |
| 32.513                 | 21                  | 63.721                 | 13                  |
| 32.697                 | 7                   | 63.911                 | 15                  |
| 32.806                 | 5                   | 66.610                 | 29                  |
| 35.586                 | 4                   | 67.009                 | 38                  |
| 38.093                 | 8                   | 67.758                 | 40                  |
| 38.631                 | 35                  | 67.921                 | 43                  |
| 43.299                 | 28                  | 69.527                 | 26                  |
| 44.316                 | 17                  | 70.475                 | 14                  |
| 44.864                 | 16                  | 71.900                 | 31                  |
| 48.007                 | 32                  | 72.610                 | 39                  |
| 48.813                 | 19                  | 72.749                 | 24                  |
| 51.531                 | 45                  | 76.407                 | 3                   |
| 53.516                 | 37                  | 79.895                 | 44                  |
| 53.776                 | 41                  | 81.165                 | 30                  |
| 54.470                 | 6                   | 83.335                 | 36                  |
| 55.589                 | 18                  | 84.200                 | 34                  |
| 58.476                 | 42                  | 88.475                 | 25                  |
| 58.936                 | 12                  | 112.335                | 23                  |
Let us display the obtained values on the numerical axis (Figure 1) and identify the taxa by the concentrations of points on it. If the points are evenly distributed, the taxa can be formed using expert valuation by identifying intervals comprising the points, which reflect the level of the potential.

![Figure 1. Projections of objects on the axis](image)

Selecting the concentrations of points on a certain interval, one can determine by table 4 those objects that fell into this interval, and then restore objects by all the indicators describing them. There will be a better picture if the amount of objects in the population undergoing the classification is large.

4. Conclusion
In conclusion, it should be noted that scientists singled out two approaches to the allocation of clusters. The first approach is based on the grouping of research objects in a multidimensional feature space, and the second is based on calculating a general indicator, for example, the potential, and grouping according to it. In the authors’ opinion, the second approach is simpler and more graphic. It can give interesting interpretations of the taxa obtained, but this is already beyond the scope of this work. The authors propose an algorithm of taxonomy on the basis of evaluation and analysis of the level of potential, reducing the problem to a one-dimensional classification.

References
[1] Sokal R R 1963 *The principles and practice of numerical taxonomy* (San Francisco: Freeman) p 359
[2] Simpson G G 1961 *Principles of animal taxonomy* (New York: Columbia University Press) p 247
[3] Sneath P H A, Sokal R R *Nature* 193 (4818) 855-860
[4] Jardine N, Sibson R 1971 *Mathematical taxonomy* (New York: Wiley) p 286
[5] Bouguettaya A, Yu Q, Liu X, Zhou X, and Song A 2015 *Expert Systems with Applications*. 42(5) 2785-2797
[6] Sarstedt M, Mooi E 2014 Cluster analysis *A concise guide to market research* (Berlin: Springer Heidelberg) pp. 273-324