UDC 519.8

O. M. Klyuchko, Ph.D., Assoc.Prof.
(National Aviation University, Kyiv, Ukraine)
T.V. Pyatchanina, Ph.D., M.G.Mazur
(Kavetsky Institute of Experimental Pathology, Oncology and Radiobiology,
National Academy of Sciences of Ukraine, Kyiv, Ukraine)

Methods of mathematics and bioinfrmatics in contemporary oncology

Brief review of wide range of cluster analysis and artificial neuron networks' methods was done for examination of their application in oncology and in linked branches of technique. A brief description of these methods basic principles that predict their use, recommendations for their application in selected areas were given.

Introduction. Contemporary publications in oncology and in linked branches of technique demonstrate the great variety of used methods of mathematics and bioinformatics. The purpose of present publication was to summarize and to classify different types of cluster analysis (CA) methods and artificial neuron networks' (ANN) methods with aim to use them in oncology, as well as in mentioned linked branches in technique. Many tasks solved in modern oncology are so complex that they require the use of complicated mathematical apparatus [1]. We have also studied series of works in which the combined application of methods of cluster analysis, artificial neural networks and image processing was demonstrated.

Review of the work done. For data analysis in oncology (and in biology, medicine as well) it is necessary sometimes to have results of hundreds or thousands of observations; the more variables are there in the problem, the more you need to have observations. In some standard computing packages (including *STNN* package) one has the means to recognize the meaningful variables, so include all variables into your examination. CA and ANN methods had proven to be effective for the solution of differential diagnosis problems in oncology. The theory of these groups of methods was observed briefly. In literature it has been demonstrated how CA and ANN methods were used for prognostic studies. It is considered that the method of neural networks is a modern approach to the personification of medical treatment and will determine the effectiveness of such radical treatment. It should be noted that because of high responsibility of work of physicians and biologists not only CA or ANN methods or even the computer diagnostics play a decisive role in the choice of methods. The purpose of such methods was to facilitate only the doctor work and to make diagnostic processes and medical treatment correct. Using of CA and ANN method permits to perfect object separation in oncology, diagnostics in medicine, and it is widely applied in contemporary practice for such purposes.

Selection of variables in CA and ANN methods are rather important task. One has to select the variables that influence on the result. With *numerical and nominal* variables in some standard computing packages (including *STNN* package) one can work directly. Variables of other types should be converted to specified

types or declared *nonessential*. If necessary, one can work with observations containin*g missed values.* Existence of data fluctuations is a difficulty. If possible, it is necessary to remove the fluctuations. If there is enough data, it is necessary to remove the missing values from the studying.

Cluster analysis (CA) is a multi-dimensional statistical procedure that collects data which contains information on object selection, then arranges them in relatively homogeneous groups. Such objects are widely studied in oncology, for example, there are great variety of cells with pathological or normal characteristics, pharmacological preparations with different types of action and other objects like these. The clustering task refers to statistical processing of such objects characteristics, as well as to a broad class of learning tasks without a teacher.

CA methods should be used to solve such problems.

1. Understanding the data by identifying of cluster structure. Splitting a sample into groups of similar objects allows you to simplify further data processing and decision making, applying to each cluster your analysis method.

2. Data compression. If the initial sample is too large, you can reduce it by leaving one of the most typical representative of each cluster.

3. Novelty detection. It is possible to find outstanding objects that can not be linked to any of the clusters.

In all these cases, hierarchical clustering can be applied, when large clusters are split into smaller ones, those for smaller ones, and etc. Such tasks are called taxonomy tasks, the result of which is a tree-like hierarchical structure. At the same time, each object is characterized by the enumeration of all clusters to which it belongs, usually from the large to the smallest.

Input data types.

Input data used for cluster analysis in oncology have to be following.

1. Description of object characteristics. Each object is described by a set of its characteristics, which are called signs. Signs can be numerical or non- numerical.

2. Matrix of distances between objects. Each object is described by distances to all other objects of metric space.

3. Matrix of similarity between objects. It is necessary to take into account the degree of similarity of an object with other sample objects in metric space. Similarity here complements the distance (difference) between objects to 1.

Example. Lets study the database of clinical analyzes. Usually it is complex one with many dimensions and different types of attributes. Attributes of these databases are divided into 2 types: numerical attributes and categorical attributes. The latter can be divided into ordered attributes and nominal attributes. For example, values of 5.29 mg, 1.6 units / liter are numerical attributes of their origin. Thus, we can determine the distance of two quantities based on the origin. Such values as "soft", "moderate", "black" belong to the attributes for which we can determine their order, but we can not determine distances in numerical values. Values such as "positive" (+), "negative" (-) are nominal attributes; They can be described, but they can not be determined by their order and distance. This means that for the analysis of such DB it is necessary to have appropriate similarity measures in order to characterize the differences between the objects in oncological practice.

Types of clustering methods classification.

A large number of clustering methods have been developed; the most widely used of them are following.

1. By models of linkage: for example, using hierarchical clustering methods, models are elaborated based on the distance of the links (connected objects).

2. Centroid models: for example, the algorithm "k-means" represents each cluster with a single vector of averages.

3. Distribution models: clusters are modeled using statistical methods of distribution, such as multivariant normal distribution using the "expectation-maximization" algorithm.

4. Density models: for example, *DBSCAN* and *OPTICS* define clusters, as connected of data space segments according to their density.

Methods of artificial neural networks: general information. Method of artificial neural networks (MANN) is widespread in science and technology and is a part of the set of widely used methods in theoretical biology and medicine. Besides of this the "simple" neural networks, which builds the system *ST Neural Networks*, are powerful weapons in the arsenal of expert in applied statistics. From the point of view of machine learning, the neural network is a partial case of clustering methods [1, 2].

Computing devices, which are assembled from simple processing elements (processors, sometimes – numerous processors) in a parallel manner, which are interacting with each other are called “Artificial Neural Networks” (ANNs). ANN is a machine that can easily be adapted to various tasks solutions. Each of such simple processor deals only with the signals that it periodically receives on the input, and the signals that it periodically output to the network constructed from other processors. Although the computational capabilities of each processor (artificial neuron - AN) are limited, however, the combination of their large numbers in a network with controlled interaction essentially increases their capabilities and they are able to solve the complex tasks. The structure of links between the network elements reflects how they are combined, and for the solution of which tasks they are designed. ANNs have a layered organization. Different layers can realize different transformations of received signals to the input of the next AN. The processed signals are shifted from the inputs of the AN on the first layer to the outputs of the last layer. ANN is a mathematical model, as well as its software or hardware implementation; it is built on the principle of the organization and operation of the networks of nerve cells of a living organism (by McCallock U., Pitts U., N. Winner, 1940s [1]). The idea of their invention is the development of methods for solving problems as the human brain does.

Characteristics f artificial neuronal networks’ methods.

1. *Neurons.* The artificial neuron (AN), which receives input signals from previous ANs, may be described by following characteristics. 1) Activation, which

depends on the discrete time parameter; 2) AN may have a threshold that remains fixed, until learning function was not changed; 3) activation function that calculates the new activation at a certain time moment, as well as new value at input; 4) output function. The output function may serve as identification function very often. The incoming AN does not have a predecessor, but this AN serves as input interface for the entire network. Accordingly, the output AN has no neuron-successor and it serves as interface for the entire network. The computing capabilities of neurons (simple processors) are usually limited by a certain rule of combining of input signals and the activation rule. This allows you to calculate the output signal basing on input signals. The output signal is transmitted to another element with a certain weighting factor, depending on the weight; the signal can either be amplified or faded.

Parallels between ANN and natural neuronal networks of the brain. Neural networks are attractive from an intuitive point of view, because they are based on the well-known biological model of the nervous systems. In the future, the development of such neuro-biological models can lead to elaboration of really “thinking computers”. Theoretical parallels between artificial neuron and natural brain neuron are in description of separate neuron characteristics.

2. *Connection and weight*. The network consists of connections (like “synapses” in natural brain networks). Each connection (predecessor) transmits a signal from input neuron to output neuron (successor). Such characteristic as “weight” has to be assigned to each connection.

3. *Distribution function.* This function calculates input value of neuron depending on previous AN outputs.

4. *Training of artificial neuron network.* Training rule is a rule or algorithm that modifies the ANN parameters in order to create a favorable exit to this network entry. This learning process usually means changing of weights and thresholds of the network. ANNs are not programmed in the usual sense of the word; they have ability to be learned. Ability to study – is one of the main advantages of ANN in comparison with traditional algorithms. Technically, during the training one has to find the coefficients of communication between the AN. In the process of learning, ANN is able to detect complex interdependencies between incoming data and output, as well as to make conclusions (generalizations). This means that in the case of successful training, the network will be able to return the correct result based on the data that was missing in the training sample (as well as in case of incomplete and / or "noisy", partially distorted data).

From the *standpoint of machine learning*, ANN is a separate case of image recognition methods, discriminate analysis, clustering methods, and so on. From a *mathematical point of view*, learning of ANN is a multi-parameter problem of

nonlinear optimization. From the *point of view of cybernetics*, ANNs are used in tasks of adaptive control and as algorithms for robotics. From the *standpoint of the development of computer technology and programming*, the neural network is a way of solving the problem of effective parallelism. From the *point of view of artificial intelligence*, ANN is the main direction of modeling of natural intelligence with the help of computer algorithms.

Conclusions

Thus, clustering methods for data analysis can be used to solve problems in oncology and linked branches of technique [1], which are related to determining whether there are two (or more) formations of the same object, or these are two different objects in terms of mathematics. Such type of problem arises in the analysis of processes of grooving of cultural masses, cell differentiation, and etc. The reason of use of cluster analysis methods is enough high for computer diagnostics for cases where it is necessary to distinguish between cells with weak differences (in normal and pathology). The new idea [1] is the application of cluster analysis methods for the solution of the task of distinguishing of elements during the elaboration of electronic databases oncology purposes.

Nowadays the use of methods of artificial neuron networks are among the most progressive practical applications in modern science. But the ANNs solve tasks not only of computer vision, speech recognition, machine translation, information filtering, and etc. With regard to biology and medicine by themselves, ANNs are used for medical diagnostics, prognosis of the course of the disease, recognition of parts of histological sections, analysis of images of computed tomography, and etc. One of such areas is computer analysis of images in oncology: for example, chromatographic studies of results of chemical and biochemical composition of samples, and etc. However, this application needs more detailed observation [1].

References

1. *Klyuchko O. M.* Information and computer technologies in biology and medicine. *Kyiv: NAU-druk.* 2008, 252 p. (In Ukrainian).