Advances in Economics, Business and Management Research, volume 139
International Conference on Economics, Management and Technologies 2020 (ICEMT 2020)

Intelligent System of Classification and Clusterization of Environmental Media for Economic Systems

A. A. Kuzmenko, L.B. Filippova*, A.S. Sazonova, R.A. Filippov

Bryansk State Technical University, Bryansk 241035, Russia
**Corresponding author. Email: libv88@mail.ru*

ABSTRACT
The article discusses the features of using Kohonen neural network in the problems of classification and clustering of environmental media. Special attention is paid to data normalization, entering mathematical relationships, and geometric interpretation.
Keywords*: threshold coefficient, Euclidian norm, centroid (semicenter), learning speed, normalization, Kohonen algorithm, self-education, stage, gradient descent*

1. INTRODUCTION

Today, the problem of classification and clustering of environmental media for meeting challenges of economic systems is particularly acute.
In the modern world, more and more attention is paid to the development of classification systems and clustering of environmental media to meet challenges of economic systems. The development of such systems allows coping with these tasks in the range of measurement, fields of scientific data analysis and interpretation. Data classification and clustering is an area that is based on the system's ability to identify features of objects from the total number and to assign objects to a specific group based on comparison of selected features or dynamically changing states of objects.
There are three stages of classifying and clustering data:
• initial processing and filtering,
* logical appraisal of filtering results,
* decision-making algorithms
This article discusses the use of Kohonen neural network as a decision-making system for environmental media classification, which is based on scientific data. Forest and herbaceous communities are considered as environmental media. Identifying of these particular media is necessary for the prediction of the possibilities of forest fire appearance and its development. It is based upon data from aerospace and satellite images.
Previously we considered the main approaches to classification and clustering of plant complexes based on various practices [5,10].

There are several ways to organize Kohonen network: self-organizing, self- correcting, and learning with a teacher. In this article, we consider the features of using a self-correcting model for classification and clustering of natural landscapes, namely herbaceous communities [1, 2, 3, 4,9].

2. NORMALIZATION OF INPUT DATA

Normalization of input data is shown in fig.1

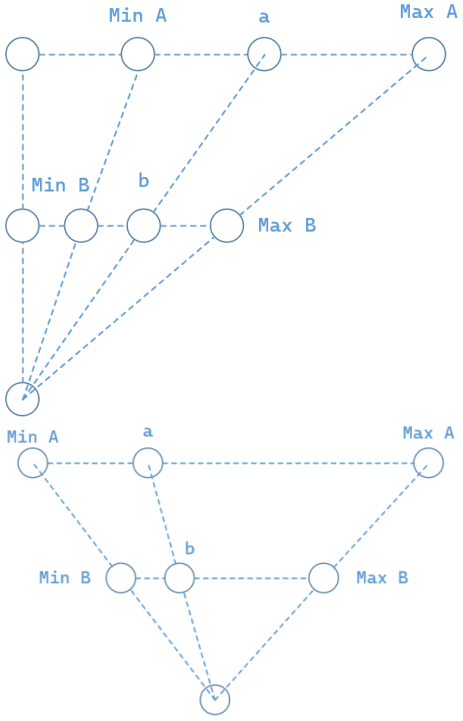

Figure 1 Normalization for number

The partial number normalization algorithm used for data approximation is shown below.

$$b = \frac{a(max\ B - min\ B)}{max\ A - min\ A} + min\ B$$

$$a = \frac{b(max\ A - min\ A)}{max\ B - min\ B} + min\ A$$

Calculate the formula for the sigmoid , where the disposal variable y takes the value in the interval - (0;1):

$$min\ B = 0, max\ B = 1$$

$$b = \frac{a(1-0)}{max\ A - min\ A} + 0 = \frac{a}{max\ A - min\ A}$$

$$a = \frac{b(max\ A - min\ A)}{1-0} + min\ A =$$

$$= b(max\ A - min\ A) + min\ A$$

For hyperbolic tangent, where the disposal variable y takes the value in the interval (-1;1):

$$min\ B = -1, max\ B = 1$$

$$b = \frac{a(1-(-1))}{max\ A - min\ A} + (-1) =$$

$$= \frac{2a}{max\ A - min\ A} - 1$$

$$a = \frac{b(max\ A - min\ A)}{1-(-1)} + min\ A =$$

$$= \frac{b(max\ A - min\ A)}{2} + min\ A$$

Fig.2 shows the main stages of normalization of the Kohonen network intended for task

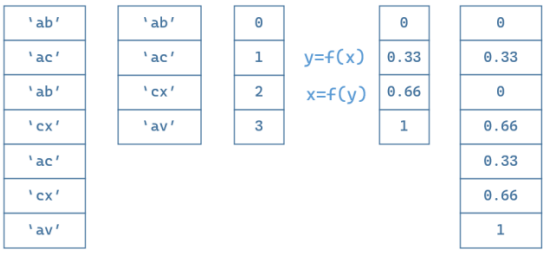

Figure 2 Stages of normalization

'ab'	0	0
'ac'	1	0.33
'cx'	2	0.66
'av'	3	1

Figure 3 Dictionary

Fig. 3 shows the idea of coding character sequence. It consists in assignment of eigenvalue to each specific sequence. The construction algorithm is outlined as follows [5]:

- all regular elements are removed, leaving only one copy.
- an eigenvalue is consequently assigned in the dictionary (in our case, an associative array in Python), then it undergoes a normalization function described above.
- The resulting dictionary can be used for either encoding the table or translating from a number to a clear human-understandable character sequence (word)

In our system for data classification the simplest encoding option is vector initializing by zeros with the only identity element. To do this, we create an array filled with zeros for each vector and assign one for the related position of the array we assign the identity element an order of magnitude optimal solution.

Aspen →0 →10000
Birch →1 →01000
Chestnut →2 →00100
Klen →3 →00010
Oak →4 →00001

Dict_network = {
'Aspen':[1, 0, 0, 0, 0],
'Birch':[0, 1, 0, 0, 0],
'Chestnut':[0, 0, 1, 0, 0],
'Klen':[0, 0, 0, 1, 0],
'Oak':[0, 0, 0, 0, 1]}

At the next stage, we make **normalization of the list using binary representation.** For this purpose we originally create a dictionary where each word stands in a certain place, i.e. it has its own number (in python, an associative array). Then this number is represented in the binary form. Since the introduction of the perceptron element of the neural network is given, the missing zeros are added to the binary number.

Aspen →0 →0_2 →00000_2
Birch →1 →1_2 →00001_2
Chestnut →2 →10_2 →00010_2
Klen →3 →11_2 →00011_2
Oak →4 →100_2 →00100_2

Besides when converting a binary number into decimal one the applicable indexes are placed over 0 and 1:

$$\overset{4\ 3\ 2\ 1\ 0}{00011_2}$$

Keep in mind that arrays start at zero and it is reasonable to store a reciprocal binary number.

Dict_network = {
'Aspen':[0, 0, 0, 0, 0],
'Birch':[1, 0, 0, 0, 0],
'Chestnut':[0, 1, 0, 0, 0],
'Klen':[1, 1, 0, 0, 0],
'Oak':[0, 0, 1, 0, 0]}

While constructing the Kohonen algorithm for self-correcting NN we should make it clear that there are three types of norm functions:

For $a = (3 \quad -12 \quad 2 \quad 4 \quad 5)$

1. Max-norm (or m-norm):

$$\|x\|_m = \max_{1 \le i \le n} |x_i|$$

$$\|a\|_m = max(3,12,2,4,5) = 12$$

2. l-norm:

$$\|x\|_l = \sum_{i=1}^{n} |x_i|$$

$$\|a\|_l = 3 + 12 + 2 + 4 + 5 = 26$$

3. Euclidian norm:

$\|x\|_b = \sqrt{\sum_{i=1}^{n} |x_i|^2}$ или $\|x\|_b^2 = \sum_{i=1}^{n} |x_i|^2$

$$\|a\|_b = \sqrt{3^2 + 12^2 + 2^2 + 4^2 + 5^2} = \sqrt{198}$$

There is a trove of:

Weight	Height	Gender
80	180	м
65	152	ж
75	164	м
60	150	ж
63	155	ж

normalizing in the range of 0 to 1

Weight	Height	Gender
1	1	м
0.25	0.074	ж
0.75	0.444	м
0	0	ж
0.15	0.185	ж
0.5	0.63	м

3. RESULTS OF COLLECTING

Let's consider an example of collecting (clustering) herbaceous communities using the implemented Kohonen algorithm. Making a point of principles of the algorithm operation we will represent all the elements in a plane, so that weight is a horizontal parameter, and height is a vertical parameter. It is quite obvious (Fig. 4) that communities of **forest vegetation** form one cumulus and communities of **herbaceous vegetation** form another one [6, 7, 8]. In our case, these cumuli are clusters. Each cluster has its own center, a so called centroid. To assign an element to one of the clusters is to find out to which centroid the element from the set is the closest.

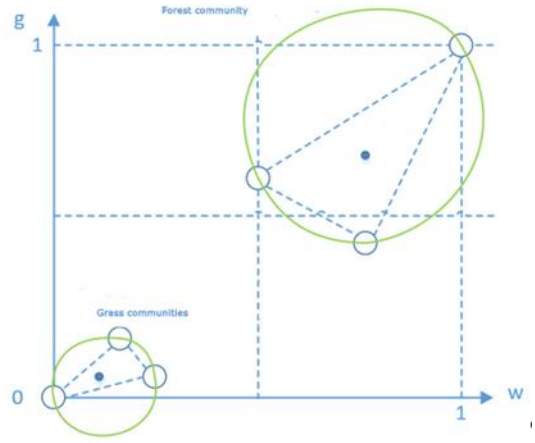

Figure 4 Kohonen algorithm (chase of cenroid))

This problem can be solved using the Pythagorean Theorem (Fig. 5):

$$R^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2$$

or the Euclidean norm for a multidimensional space*:*

$$R_j = \sqrt{\sum_{i=1}^{M} (\bar{x}_i - w_{ij})^2}$$

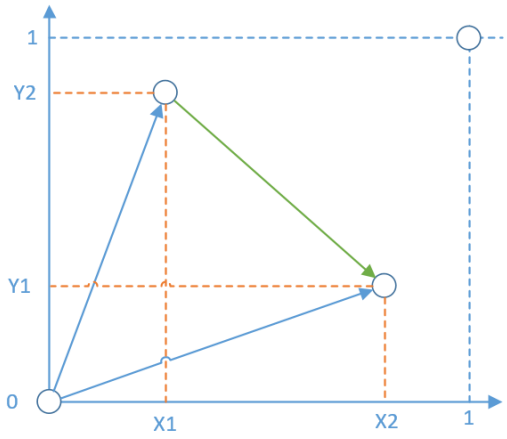

Figure 5 Calculating the distance between two points

Besides making a scalar product of vectors we get the largest number showing to which centroid the element from the set is the closest..

Let's consider the features of initializing weights:

[-1;1]:$|w|_{ij} \le \frac{1}{\sqrt{M}}$, where M – dimension figure of entry.

[0; 1]:$0.5 - \frac{1}{\sqrt{M}} \le w_{ij} \le 0.5 + \frac{1}{\sqrt{M}}$

After random generating of neural network weights, i.e. fortuituos placing of a point in the space in the tolerance region relative to normalization (normalized numbers in the space from 0 to 1 by x and by y), you need to answer the question: how to change the coordinates of the centroid so that it becomes the center of a cluster of communities. For answering this question, it is enough to find out the difference between the centroid and the element of the set vertically and horizontally, then take a small portion of this difference, i.e. the learning speed, and change the position for this portion. In other words we should calculate "gradient descent".

Let's introduce notations and functions.

Let's introduce notation and functions

X – an assembly vector, having *x, y- coordinates*

W – centroid having coordinates *w1*, *w2* in x,y-direction

Let's activize net-

$$y_j = X * W_j = \sum_{i=1}^{m} w_{ji}\, x_i$$

we seek the maximal element

$$j_{\max} = \arg\max_j \{y_j\}$$

Move the centroid to the winner:

$$w_{ij}(n+1) = w_{ij}(n) + v(\bar{x}_i - w_{ij}^{(q)})$$

where $q = j_{max}$, v – lerning speed

It is worth moving the centroid until there has been a significant change in the weight coefficients within the specified accuracy during the last stage of training.
Let's take a different set of parameters of herbaceous communities (data in the example are modeled, while numerical indicators are taken at random) and normalize it:

Number of different kinds of	General projective cover	Community type
1	1	1
0	0.25	0
0.75	0.6	1
0.4	0.4	0
0.25	0	0
0.5	0.75	1
1	0.8	1
0.2	0.25	0

In this connection the community type: 1- Forest, 0 – Herbaceous
Three examples of how the centroid moves to the center of the cluster (Fig. 6). Red is given through weights for orange (**forest community),** while green is intended for blue (**herbaceous community**).

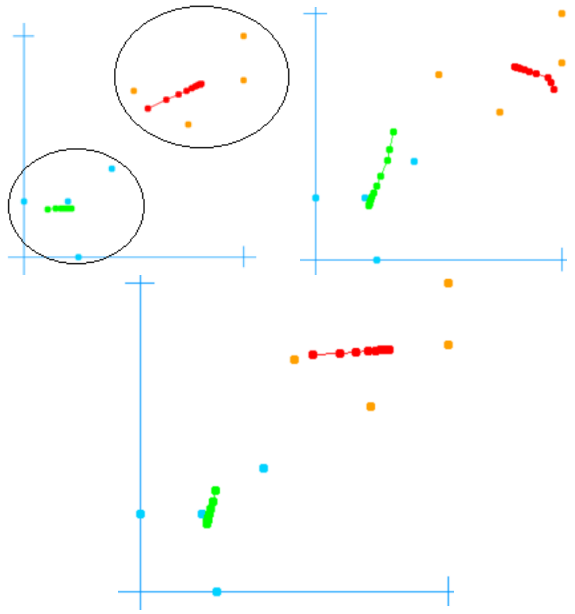

Figure 6 Centroid motion relative to the cluster centers.

4. CONCLUSION

The article considers two closely related technologies: normalization for both: numbers and a set of sequences (words) and construction of a self-correcting Kohonen network. The network divides plant communities into 2 types: forest and herbaceous according to the specified parameters. This system allows you to divide plant communities not only into two groups. The approach described in the article is aimed at simplifying the perception of information. The overall accuracy rate was 98% in the assay test, which is a high indicator for this task. For now the developed system makes ecological and floristic classification of vegetation identifying communities based on 27 features possible.

REFERENCES

[1] Barsky, A. B. Logical neural networks: a Textbook, Moscow: Binom, 2013, 352 p.
[2] Galushkin, A. I. Neural networks: fundamentals of the theory, Moscow: RandS, 2014, 496 p.
[3] Galushkin, A. I. Neural networks: the history of theory development: a textbook for universities, Moscow: Alliance, 2015, 840 p.
[4] Redko, V. G. Evolution, neural networks, intelligence: Models and concepts of evolutionary Cybernetics. Moscow: Lenand, 2015, 224 p.
[5] Kuzmenko A. A., Kondrashin D. E., Methods and approaches to the development of a system for automated analysis of the dynamics of changes in the area of forest stands based on pattern automatic recognition methods, Bryansk, ERGODESIGN Vol. 2019 № 4 (6), 230-240.
[6] Hughes G.F. On the mean accuracy of statistical pattern recognizers, IEEE Transactions on Information Theory, 1968, IT-14, P. 55-63.
[7] Devroye L. Gyorfi L., Lugosi G.A. Probabilistic Theory of Pattern Recognition, Berlin: Springer–Verlag. 1996.
[8] Adams J.B., Smith M.O., Johnson P.E. Spectral mixture modeling: A new analysis of rock and soil types at the Viking Lander 1 site, Journal of Geophysical Research, 1986, V. 91, P. 8098-8112.
[9] E.A. Leonov, Y.A. Leonov, Y.M. Kazakov, L.B. Filippova, Intellectual subsystems for collecting information from the internet to create knowledge bases for self-learning systems, In: Abraham A., Kovalev S., Tarassov V., Snasel V., Vasileva M., Sukhanov A. (eds), electronic, Proceedings of the Second International Scientific Conference "Intelligent Information Technologies for Industry" (IITI'17). IITI 2017. Advances in Intelligent Systems and Computing, 2017, vol 679, Springer, Cham, p. 95-103, DOI:10.1007/978-3-319-68321-8_10
[10] YU.A. Leonov, E.A. Leonov, A.A. Kuzmenko, A.A. Martynenko , E.E. Averchenkova, R.A. Filippov Selection of rational schemes automation based on working synthesis instruments for technological processes, Yelm, WA, USA: Science Book Publishing House LLC, 2019, 192 p, ISBN: 978-5-9765-4023-1.