I.J. Modern Education and Computer Science, **2025, 4, 101-111**
Published Online on August 8, 2025 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2025.04.07

Fuzzy Clustering of Educational Data with Automated Selection of Processing Parameters in System Analysis of Quality Education

Zhengbing Hu
School of Computer Science, Hubei University of Technology, Wuhan, China
E-mail: drzbhu@gmail.com
ORCID iD: https://orcid.org/0000-0002-6140-3351

Oleksandr Derevyanchuk
Yuriy Fedkovych Chernivtsi National University, Chernivtsi, 58012, Ukraine
E-mail: o.v.derevyanchuk@chnu.edu.ua
ORCID iD: https://orcid.org/0000-0002-3749-9998

Serhiy Balovsyak
Yuriy Fedkovych Chernivtsi National University, Chernivtsi, 58012, Ukraine
E-mail: s.balovsyak@chnu.edu.ua
ORCID iD: https://orcid.org/0000-0002-3253-9006

Yuriy Ushenko*
Yuriy Fedkovych Chernivtsi National University, Chernivtsi, 58012, Ukraine
E-mail: y.ushenko@chnu.edu.ua
ORCID iD: https://orcid.org/0000-0003-1767-1882
*Corresponding Author

Hanna Kravchenko
High State Educational Establishment «Chernivtsi transport college», Chernivtsi, 58000, Ukraine
E-mail: hannakravchenko81@gmail.com
ORCID iD: https://orcid.org/0009-0004-7609-0345

Iryna Sapsai
Institute of Postgraduate Education, Borys Grinchenko Kyiv Metropolitan University, Kyiv, 02152, Ukraine
E-mail: i.sapsai@kubg.edu.ua
ORCID iD: https://orcid.org/0000-0002-7338-715X

Received: 18 April, 2025; Revised: 26 May, 2025; Accepted: 27 June, 2025; Published: 08 August, 2025

Abstract: Clustering of educational data was performed in the space of two parameters using the K-Means method. Students who are characterized by grades in certain types of activities were used as objects of clustering. Software for fuzzy data clustering is implemented in the Python language in the Google Colab cloud service. The obtained clusters are described by fuzzy Gaussian membership functions, which allowed to reliably determine the membership of each object to a certain cluster, even if the clusters do not have clear boundaries. Due to clustering, the most important characteristics of the educational process for a certain task are obtained, that is, this is how Data Manning tasks are solved. Fuzzy membership functions implemented using the scikit-fuzzy library. The developed program can also be used for educational purposes, as it allows a better understanding of the principles of cluster analysis and fuzzy logic. The correctness of the work of the developed program was confirmed during the processing of test educational data. The determination of the number of clusters was performed by software, taking into account the intra-cluster and inter-cluster distances, as well as the shape of the clusters. Automated selection of the number of clusters and cluster boundaries allows to reduce data processing time. The developed clustering tools are designed to increase the efficiency of system analysis of quality education.

Index Terms: Education Technology, Clustering methods, Data Mining, Educational Data, Fuzzy Logic, K-Means, System Analysis, Quality Education

1. Introduction

Currently, Data Mining methods are widely used for processing educational data, since the volume of such data is constantly increasing and their manual processing is time-consuming [1, 2, 3]. For example, a common task is the analysis of student grades in certain subjects. Due to the methods of Data Manning, it is possible to automatically process large volumes of data of various types [4, 5, 6, 7, 8]. As a result of clustering, the initial set of objects (data) of a large size is divided into subsets (clusters), which allows to structure the data and purposefully process the objects of individual clusters. In particular, students or study groups are considered as objects of clustering. Each clustering object is described by a number of parameters (characteristics): grades in certain subjects, number of points for test tasks, etc. For example, when clustering students' learning outcomes, students' grades are used as indicators for clustering. In most cases, clustering is performed on the basis of two features (parameters), but clustering is also possible in the space of a larger number of parameters. A number of clustering algorithms are used, in particular, Hierarchy Algorithms, Agglomerative Nesting, Divisive Analysis [9]. Hierarchical methods include BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies), which forms a tree-like structure and is effective for processing large data sets. That is, in the BIRCH method, small-sized clusters are processed at low levels of the hierarchy, and at high levels, such clusters are combined with each other. The DBSCAN method (Density-based spatial clustering of applications with noise) is effective under the condition that the clusters form dense (compact) areas in the space of their characteristics. Such methods can be used when processing data with noise [10].

A common clustering method is the K-means method, the advantages of which are simplicity of software implementation and high speed [11]. In addition, the advantage of the K-means method is its low computational (time) complexity (relative to similar methods), which is O(N), where N is the number of objects under study [12, 13]. In comparison, for the DBSCAN method, the computational complexity is not less than O(N log$N$), and the computational complexity of hierarchical clustering is not less than O(N^2 log$N$) [14]. The low computational complexity of the K-means method allows for high-speed clustering of even large volumes of data (N > 1000), i.e. the K-means method scales well for processing large data sets. Therefore, the work uses the K-Means method for data clustering, which divides the initial data into non-overlapping clusters. However, in the K-means method there is a problem of choosing the quantity of clusters. It is possible to automate the selection of the quantity of clusters, in particular, taking into account intra-cluster and inter-cluster distances. When determining the quantity of clusters, the conditions of a specific task are also taken into account; for example, if students need to be divided into three groups based on the results of their studies, then the quantity of clusters is chosen equal to the quantity of such groups. However, in most cases, the quantity of clusters needs to be determined. Cluster boundaries are described by line segments, circles, ellipses, or arcs, so it is advisable to automate the selection of cluster boundaries as well.

Cluster analysis is widely used for processing educational data, which is described in works [15, 16, 17]. Clustering methods are used to analyze of student behavioral patterns in work [15], which allows for purposeful work with different groups of students. It is shown that for the same data it is advisable to use different clustering methods, for example hierarchical methods, DBSCAN and k-means. From the obtained clustering results, the best ones are obtained, for example, according to the criteria of intra-cluster and inter-cluster distances. The paper [16] describes how the assessment of student learning outcomes is performed using clustering methods in modern educational platforms. Work [17] shows that by means of clustering of educational data it is possible to analyze and correct students' learning in electronic systems, to predict their achievements. Due to clustering, it is possible to investigate complex relationships between data of various types [18], in particular, between educational data. Work [19] describes methods of adapting the educational process to the needs of each student, the advantages of taking into account the peculiarities of student learning. The application of intellectual data analysis contributes to the improvement of educational achievements of students [20, 21, 22]. In works [23, 24, 25], the possibilities of using deep learning methods and artificial neural networks to evaluate the effectiveness of the educational process were investigated. The reviewed works show that modern clustering methods perform automatic analysis of significant volumes of educational data according to various parameters, which increases the accuracy and speed of system analysis of the quality education.

In the simplest cases, the clusters are compact and clearly separated. However, when processing educational data, clusters often overlap, which leads to the problem of separating objects at cluster boundaries. For example, if a certain student is close to the boundary of two clusters, then taking into account the student's belonging to only one cluster will lead to the loss of information about his partial belonging to the other cluster. Even a slight change in the parameter of an object located on the boundary can lead to his belonging to the other cluster. If we consider clusters as fuzzy sets, then it is possible to describe the belonging of objects on the boundary simultaneously to several clusters by fuzzy membership functions [26, 27, 28, 29]. Fuzzy logic capabilities are used to process data of various types, for example, when segmenting images. The type of fuzzy membership function should be chosen taking into account the features of the cluster shapes, which are characteristic of the studied data. Therefore, the purpose of this work is to increase the

accuracy of clustering of educational data by means of fuzzy logic [26], as well as to reduce the complexity of data processing due to the automatic determination of the number and boundaries of clusters. The topic of the work is relevant, since the analysis of the results of students' learning by means of Data Manning allows improving the educational process [30].

2. Theoretical Foundations of Data Clustering Using Fuzzy Logic

2.1. Principles of data clustering

Educational data contains information about N objects, each of which is characterized by several parameters. We will perform the clustering process on the basis of 2 parameters of objects (in the space of 2 features), therefore, from all parameters, two parameters x_1 and x_2 are selected, which are the most important for a certain task [31]. Such parameters can mean, for example, the grades of N students in two subjects. The x_1 and x_2 parameters of the objects correspond to the x_1 and x_2 axes of the rectangular Cartesian coordinate system, so the data objects under study are visualized as points with coordinates (x_1, x_2).

Mathematically, objects are clustered based on their parameters $x_1(i)$ and $x_2(i)$, where i = 1,..., N. The distance $\rho(i, m)$ between objects with numbers i and m is calculated as the Euclidean distance or distance of city blocks (Manhattan distance). In most cases, the Euclidean distance is used. Quantification of clustering quality is performed on the basis of intra-cluster distance (1) and inter-cluster distance (2) [12-14]. For a clear separation of clusters (with minimal overlap), the average intra-cluster D_{IN} distance should be minimal:

$$D_{IN} = \frac{\sum_{i=1}^{N}\sum_{m=1}^{N}[c_i=c_m]\cdot\rho(\mathrm{i,m})}{\sum_{i=1}^{N}\sum_{m=1}^{N}[c_i=c_m]} \to min, \quad (1)$$

where c_i is the cluster number for the object with number i; c_m is the cluster number for the object with number m.

For better separation of clusters, the average inter-cluster distance D_{OUT} should be maximal:

$$D_{OUT} = \frac{\sum_{i=1}^{N}\sum_{m=1}^{N}[c_i\neq c_m]\cdot\rho(\mathrm{i,m})}{\sum_{i=1}^{N}\sum_{m=1}^{N}[c_i\neq c_m]} \to max. \quad (2)$$

The analysis of the quality of the clustering results is simplified if we use the ratio D_R (3) of intra-cluster D_{IN} distances to inter-cluster distances D_{OUT}:

$$D_R = D_{IN}/D_{OUT} \to min. \quad (3)$$

The distance ratio D_R depends on the clustering method used and on the quantity of clusters. A smaller D_R value means higher reliability of clustering results. However, the analysis of distances D_R does not take into account the shape of the clusters. Compact clusters with a shape close to a sphere are convenient for processing. Therefore, the study of the shape of clusters was performed by analyzing their average eccentricity value E_C [32], which describes the symmetric or asymmetric distribution of objects relative to the center of the cluster. To calculate the eccentricity E_C, the centers of gravity of the C_{x1} and C_{x2} clusters are first calculated relative to the x_1 and x_2 coordinate axes. Discrete central moments of the cluster are calculated according to the formulas:

$$\mu_{11} = \frac{1}{C_q}\cdot\sum_{i=1}^{N}(x_1(i) - C_{x1})\cdot(x_2(i) - C_{x2}), \quad (4)$$

$$\mu_{20} = \frac{1}{C_q}\cdot\sum_{i=1}^{N}(x_1(i) - C_{x1})^2, \quad (5)$$

$$\mu_{02} = \frac{1}{C_q}\cdot\sum_{i=1}^{N}(x_2(i) - C_{x2})^2, \quad (6)$$

where C_q is the quantity of objects in the cluster.

Eccentricity E_C is calculated through discrete central moments according to the formula:

$$E_C = \frac{(\mu_{20}-\mu_{02})^2+4\mu_{11}^2}{(\mu_{20}+\mu_{02})^2}. \quad (7)$$

If the shape of the cluster has a clear orientation, then $E_C \approx 1$; if the shape of the cluster does not have a certain orientation, then $E_C \approx 0$; in intermediate cases $E_C \approx 0.5$.

2.2. Using fuzzy membership functions to describe clusters

Fuzzy membership functions, for example, triangular or Gaussian, are used to determine whether a certain object belongs to clusters [26, 27]. Triangular functions are simple to implement, but Gaussian functions provide a more correct description of the belonging of objects to clusters, since they change smoothly. Mathematically, fuzzy Gaussian membership function, which depends on the x coordinate of objects, is described by the formula:

$$\mu(x, x_0, \sigma) = exp\left(\frac{-(x-x_0)^2}{2\sigma^2}\right), \quad (8)$$

where x_0 is the center of the function; σ is the standard deviation.

As a result of clustering, N objects are divided into Q_k clusters with numbers $k = 1,..., Q_k$. Since during clustering, each object is described by coordinates x_1 and x_2, therefore we will introduce fuzzy membership functions $\mu_{x1}(k, d_{x1})$ and $\mu_{x2}(k, d_{x2})$. The fuzzy function $\mu_{x1}(k, d_{x1})$ describes the extent to which an object belongs to cluster # k, where d_{x1} is the distance of the object to the center of cluster # k by coordinate x_1. Similarly, the fuzzy membership function $\mu_{x2}(k, d_{x2})$ describes the degree to which an object belongs to cluster # k, where d_{x2} is the distance of the object to the center of cluster # k by coordinate x_2. The fuzzy membership function $\mu_\rho(k, \rho)$ describes the extent to which an object belongs to cluster # k, where ρ is the distance of the object to the center of cluster # k by coordinates x_1 and x_2.

In the case of compact and symmetric clusters, it is convenient to describe their boundaries with circles (with radius R_c) or ellipses. However, when processing educational data, the shapes of clusters are often asymmetric, so it is advisable to describe the boundaries of such clusters with arcs of circles with radii:

- R_{cL} is the radius of the left arc for the cluster boundary.
- R_{cR} is the radius of the right arc for the cluster boundary.
- R_{cDn} is the lower arc radius for the cluster boundary.
- R_{cUp} is the radius of the upper arc for the cluster boundary.

For the studied objects, their fuzzy functions of belonging to clusters with numbers $k = 1,..., Q_k$ are calculated. The measure of object belonging to the cluster with number k is calculated as the value of the fuzzy membership function $\mu_\rho(k, \rho)$ according to the formula:

$$\mu_\rho(k, \rho) = 1 - \frac{\rho}{R_c \cdot k_R}, \quad (9)$$

where ρ is the distance of the object to the cluster center with number k (by coordinates x_1, x_2);

R_c is the cluster radius # k; k_R is the cluster size factor (for example, $k_R = 1.5$)

The cluster size factor k_R is used to adjust the overlap of clusters (if there is significant data correlation, there will be a large overlap of clusters and the value of k_R should be increased). The values of the membership function lie in the range from 0 to 1. For example, if the distance ρ of the object to the center of the cluster with the number k is equal to 0, then $\mu_\rho(k, \rho) = 1$, that is, the object completely belongs to the cluster with the number k. If the distance ρ is equal to or greater than $(R_c \cdot k_R)$, then $\mu_\rho(k, \rho) = 0$, that is, the object does not belong to the cluster with number k at all.

Fuzzy membership functions are also calculated, which describe the belonging of an object to a cluster taking into account its coordinates:

- $\mu_{x1}(k, d_{x1})$ is a fuzzy function of the object belonging to cluster # k, where d_{x1} is the distance of the object to the center of the cluster by coordinate x_1.
- $\mu_{x2}(k, d_{x2})$ is a fuzzy function of object belonging to cluster # k, where d_{x2} is the distance of the object to the center of the cluster by coordinate x_2.

The fuzzy function of belonging $\mu_{x12}(k, d_{x1}, d_{x2})$ of an object to cluster # k by coordinates d_{x1}, d_{x2} is calculated by the formula:

$$\mu_{x12}(k, d_{x1}, d_{x2}) = \frac{\sqrt{(\mu_{x1}(k,d_{x1}))^2 + (\mu_{x2}(k,d_{x2}))^2}}{\sqrt{2}}, \quad (10)$$

The values of the fuzzy membership functions $\mu_\rho(k, \rho)$, $\mu_{x12}(k, d_{x1}, d_{x2})$, $\mu_{x1}(k, d_{x1})$ and $\mu_{x2}(k, d_{x2})$ can be specified by changing the cluster size factor k_R, for example, for refinement of the function $\mu_\rho(k, \rho)$ in formula (9). This allows for more accurate cluster analysis for educational data.

3. Software Implementation

Software implementation of fuzzy clustering of educational data is made in Python using Google Colab cloud service and Jupyter Notebook [33, 34]. In the process of clustering, N objects under study are divided into Q_k clusters based on the parameters of objects x_1 and x_2.

The developed program first reads the initial educational data (for example, student grades in certain subjects), from which two parameters (x_1 and x_2) are extracted. Next, the quantity of clusters Q_k is set either manually, or the program cycles through the values of Q_k in the set range. For each Q_k the average intra-cluster distance D_{IN}, the average inter-cluster distance D_{OUT} and the distance ratio D_R are calculated according to formulas (1-3). Based on the distance ratio D_R, a specific value of the quantity of clusters Q_k is selected. Software determination of the quantity of clusters reduces the subjective factors associated with the work of the performer.

The initial data are recorded in arrays $x_1(i)$ and $x_2(i)$, where $i = 1,..., N$. Data clustering is performed by the k-means method, which is implemented in the KMeans(Q_k) function of the Sklearn library [35]. As a result of clustering, the number of the cluster $k = 1,..., Q_k$ to which it belongs is determined for each object. The defined cluster numbers for the objects are stored in the array $N_k(i)$, where $i = 1,..., N$. The coordinates x_1 and x_2 of their centers of gravity are calculated for the obtained clusters, which are stored in the arrays $C_{x1}(k)$ and $C_{x2}(k)$ respectively.

For each cluster k, the quantity of objects C_q, the radii of the cluster boundary arcs R_c, R_{cL}, R_{cR}, R_{cDn}, R_{cUp}, as well as the fuzzy membership functions $\mu_p(k, \rho)$, $\mu_{x1}(k, d_{x1})$, $\mu_{x2}(k, d_{x2})$, $\mu_{x12}(k, d_{x1}, d_{x2})$ of the object to the cluster are calculated. Gaussian fuzzy membership functions are calculated by the gaussmf() function of the scikit-fuzzy library. A characteristic feature of Gaussian membership functions is a smooth change in their values (according to the normal law), so the belonging of objects to clusters is determined correctly.

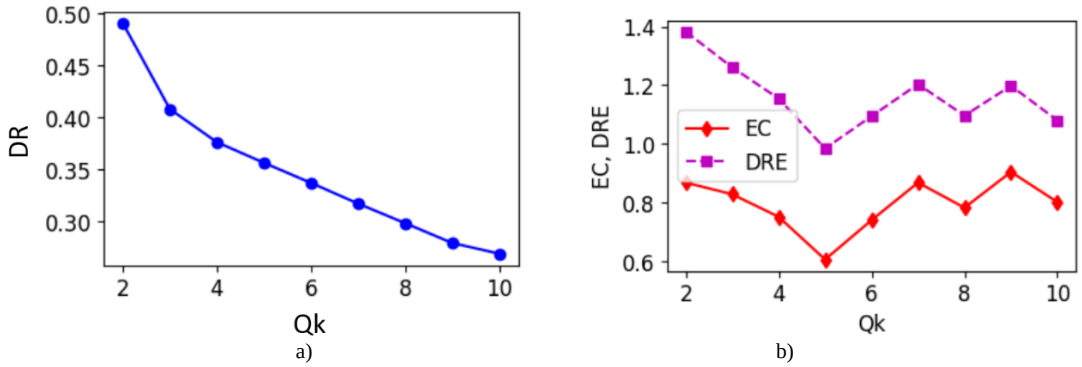
Fuzzy membership functions make it possible to calculate the degree of belonging of the studied objects to each cluster. After that, the obtained clusters, their centers, boundaries and fuzzy membership functions are visualized. Clusters with a symmetrical distribution of objects relative to the center can be distinguished by straight line segments, circles or ellipses. However, for clusters of asymmetric shape (this shape of clusters is often found in educational data processing), it is more appropriate to describe the cluster boundaries by arcs for each quadrant. This allows for more accurate spatial localization of clusters, which simplifies the visual perception of clustering results for their further analysis. Cluster boundaries are described by arcs with radii R_{cL}, R_{cR}, R_{cDn}, R_{cUp}. The values of arc radii are calculated by the coordinate descent method (with a minimum step, for example, 1) under the condition that the radii are minimal and all objects are inside the arcs of the corresponding cluster.

4. Results and Discussion

By developed program the clustering and analysis of test educational data [36] was performed, which contained the results of student learning (grades from several subjects, the values of which are in the range from 0 to 100). Data preprocessing consisted of discarding objects with missing values and outliers (anomalies) (objects whose parameter values are outside the allowable ranges), as well as normalizing clustering parameters (scaling to one range). Normalization ensures that all object parameters are taken into account equally when clustering them. At the user's discretion, data preprocessing is possible, in which objects with missing values and outliers are not removed, and their parameter values are replaced with average or interpolated values. Clustering was performed according to the "math score" (x_1-coordinate during clustering) and "writing score" (x_2-coordinate during clustering) parameters. Each of the parameters x_1 and x_2 takes N values ($N = 1000$). Clustering using other parameters is performed similarly.

Using the K-Means method, data clustering was performed with the quantity of clusters Q_k, which varied in the range from 2 to 10. For each Q_k, the average intra-cluster distance D_{IN}, the average inter-cluster distance D_{OUT}, and the distance ratio D_R were calculated (Fig. 1a). Calculations were made using Euclidean and Mathetian distances. In both cases, similar results were obtained, confirming the stability of clustering. In the following, Euclidean distance was used.

Based on the distance ratio D_R, it is advisable to choose the largest value of the quantity of clusters Q_k =10. An analysis of the shape of the clusters, which is described by the E_C eccentricity (7) was also carried out (Fig. 1b). To select the optimal quantity of clusters Q_k, the minimum refined distance ratio $D_{RE} = D_R + E_C$ was used. The clustering process was performed 10 times, as a result of which the minimum D_{RE} value was obtained for $Q_k = 5$ (Fig. 1b). In addition, when $Q_k > 5$, the analysis of the obtained clusters is complicated, since in this case it is more difficult to describe the characteristics of students who, according to their assessments, entered a certain cluster. Therefore, for further clustering, the value Q_k =5 was chosen, which provides a better separation of clusters (compared to smaller values of Q_k).



a) b)

Fig. 1. Graph of the dependence of the distance ratio DR (a), eccentricity EC and refined distance ratio DRE (b) on the quantity of clusters Qk

By the help of developed software using the K-Means method obtained the following results (Fig. 2).

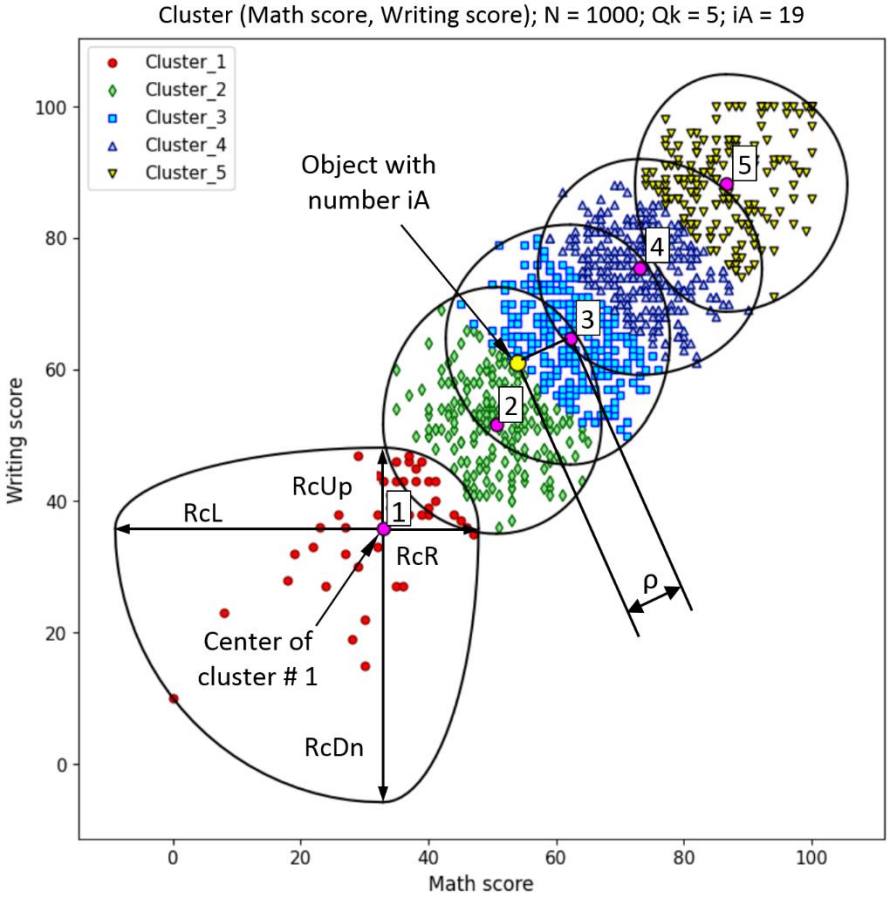


Fig. 2. The result of clustering test educational data according to the parameters "math score", "writing score"; cluster boundaries are limited by arcs with radii RcL, RcR, RcDn, RcUp

Each cluster with the number k corresponds to a set of students who are characterized by certain learning outcomes:

1 – low grades in mathematics and writing.
2 – higher than low grades in mathematics and writing.
3 – lower than average grades in mathematics and writing.
4 – average grades in mathematics and writing.
5 – high grades in mathematics and writing.

Based on the obtained clusters, it is possible to study the results of the educational process. The program allows to analyze the study results of a student with an i_A number. For such a student, the distances to the centers of clusters with numbers k, where $k = 1,...,Q_k$ are calculated: distance d_{x10} along the x coordinate, distance d_{x20} along the x_2 coordinate, Euclidean distance ρ_0. Based on the calculated distances d_{x10}, d_{x20}, ρ_0 the degree of student belonging to cluster k is determined as the value of the fuzzy membership functions $\mu_{x1}(k, d_{x10})$ (Fig. 3), $\mu_{x2}(k, d_{x20})$ (Fig. 4), $\mu_{\rho}(k, \rho_0)$ (Fig. 5).

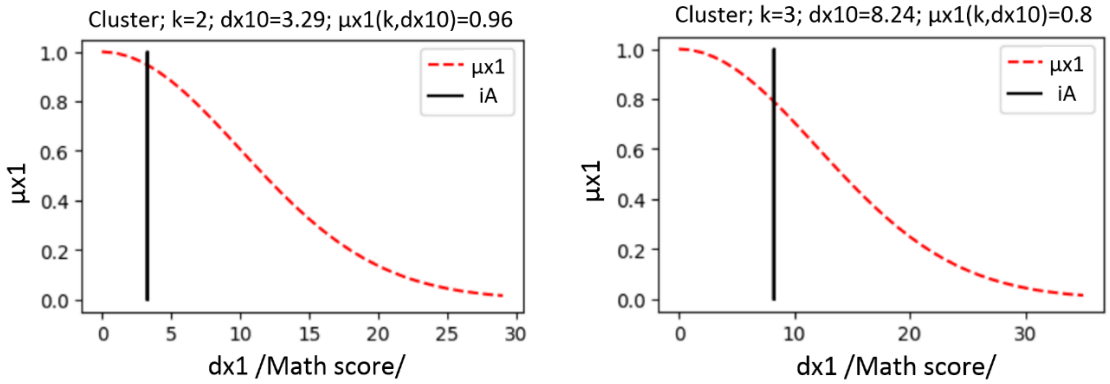


Fig. 3. Fuzzy functions of belonging μx1(k, dx1) of students to clusters with numbers k = 2, 3 according to "math score"; the vertical segment shows the value of the distance dx10 for the analyzed student with number iA =19 (Fig. 2)

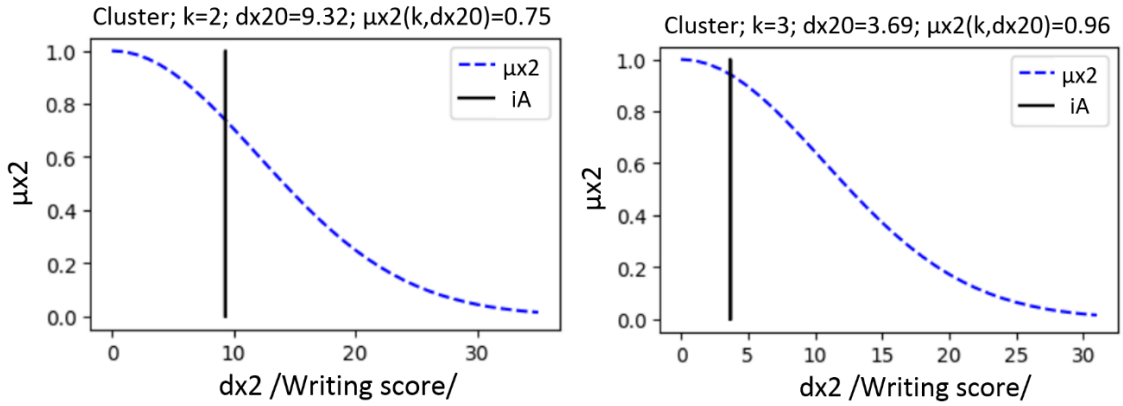


Fig. 4. Fuzzy functions of belonging μx2(k, dx2) of students to clusters with numbers k = 2, 3 according to "writing score"; the vertical segment shows the value of the distance dx20 for a student with number iA =19 (Fig. 2)

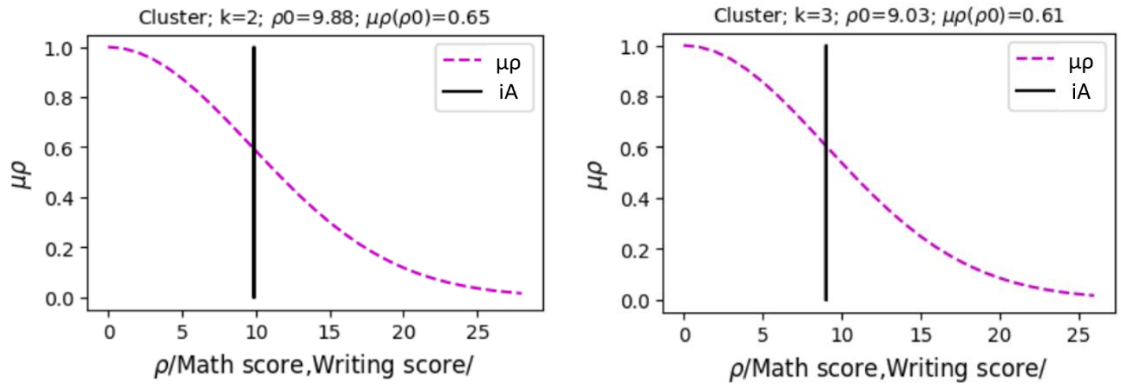


Fig. 5. Fuzzy functions of belonging μρ(k, ρ) of students to clusters with numbers k = 2, 3 based on "math score" and "reading score"; the vertical segment shows the value of the distance ρ0 for the student with number iA = 19 (Fig. 2)

With the use of fuzzy membership functions $\mu_{x1}(k, d_{x10})$ (Fig. 3) and $\mu_{x2}(k, d_{x20})$ (Fig. 4), the values of degrees of membership of the analyzed student with number i_A to all clusters with numbers k (Fig. 6) were calculated as functions $\mu_{x1R}(k)$ and $\mu_{x2R}(k)$, respectively. With the use of fuzzy functions $\mu_{\rho}(k, \rho_0)$ (Fig. 5), the value of the degree of belonging of the analyzed student with number i_A to clusters with numbers k (Fig. 7) was calculated as the value of $\mu_{\rho R}(k)$. The value of the degree of belonging of the student with number i_A to clusters with numbers k is calculated as the value $\mu_{x12R}(k)$ according to formula (6), using the values $\mu_{x1R}(k)$ and $\mu_{x2R}(k)$. Since the asymmetric form of clusters is taken into account when calculating $\mu_{x12R}(k)$, the values of $\mu_{x12R}(k)$ more correctly show the belonging of students to clusters (compared to $\mu_{\rho R}(k)$).

Due to this, the values of $\mu_{xyR}(k)$ allow accurate assessment and analysis of the student's educational achievements. For example, the analyzed student with the number i_A (Fig. 7) mostly belongs to cluster 3 ($\mu_{x12R}(3) = 0.884$) is the fuzzy set "below average math and writing scores", and to a lesser extent belongs to cluster 2 ($\mu_{x12}(2) = 0.861$) is the fuzzy set "above low math and writing scores".

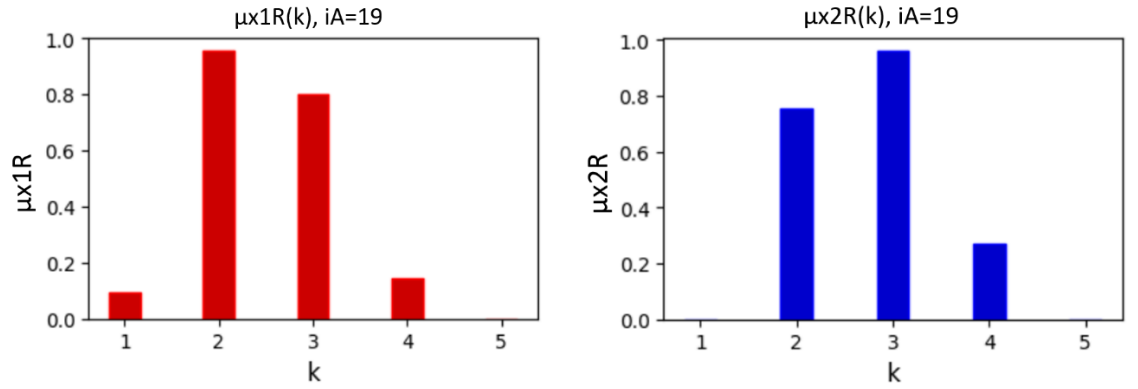


Fig. 6. The value of the degree of belonging of the analyzed student with number iA =19 (Fig. 2) to clusters with numbers k, calculated on the basis of the membership functions μx1(k, dx10) (Fig. 3) and μx2(k, dx20) (Fig. 4)

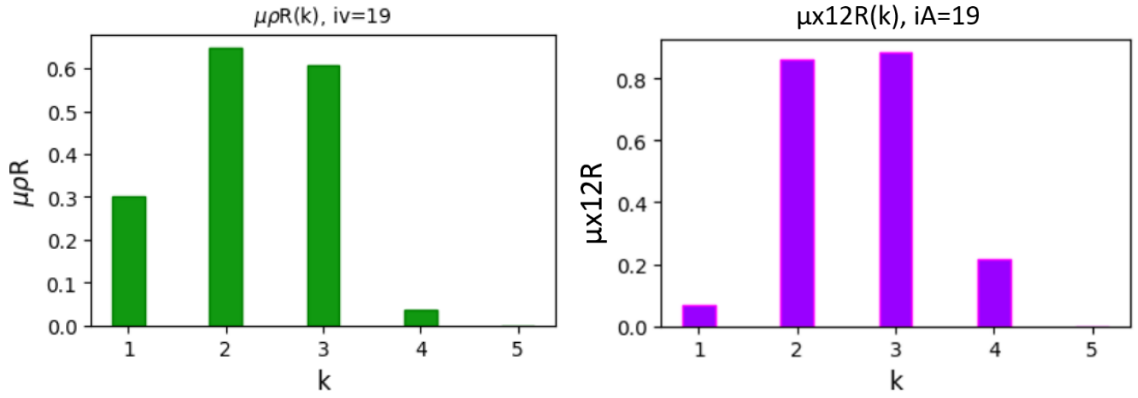


Fig. 7. The value of the degree of belonging of the analyzed student with number iA =19 (Fig. 2) to clusters with numbers k, calculated on the basis of fuzzy membership functions μρ(k, dρ0) (Fig. 5) and μx12(k, dx10, dx20) (10)

Each obtained cluster contains students with a certain level of competence, so such clustering results can be used in the educational process to determine the level of difficulty of tasks (tests, laboratory work). In real cases, students may partially belong to different clusters (which overlap), so such students are offered tasks of different levels of difficulty. For example, the analyzed student with number i_A is offered tasks with difficulty level 3 and additionally tasks with level 2. The obtained clustering results are intended to increase the efficiency of the system analysis of the quality education, which involves targeted data collection, processing and adjustment of the educational process. The division of students into clusters according to their educational achievements increases the accuracy of monitoring learning outcomes. Based on the results of such monitoring, correction of individual educational trajectories of students or elements of academic disciplines is performed [37, 38]. Cluster analysis is effective in processing educational data of various types: scores in academic disciplines, test results [39] and learning styles [40, 41], etc.

5. Conclusions

Software tools for fuzzy clustering of educational data have been developed, which is an integral part of the system analysis of the quality education. The clustering program was implemented in Python on the Google Colab cloud service. The developed program uses the scikit-learn library for clustering, and the scikit-fuzzy library for working with fuzzy functions. Clustering was performed using the K-Means method in the space of two features, which were used to evaluate students in certain subjects. Automated determination of the number of clusters and their boundaries based on intra-cluster and inter-cluster distances made it possible to minimize subjective factors and reduce data processing time.

The novelty of the work is the description of objects belonging to clusters by their own fuzzy Gaussian functions μ_p, μ_{x1}, μ_{x2} and μ_{x12}, which makes it possible to correctly determine the belonging of objects to several clusters.

Testing of the developed program when processing real educational data showed correct results. The use of fuzzy membership functions made it possible to correctly calculate the degree of belonging of the studied objects to several clusters, even if the objects are located on the borders of the clusters. Cluster analysis of students' educational achievements allows to individually determine the recommended level of difficulty of tasks for each student. The developed program can also be used for educational purposes when studying the principles of clustering and fuzzy logic.

References

[1] R. Ahuja, A. Jha, R. Maurya and R. Srivastava, *Analysis of Educational Data Mining. In Harmony Search and Nature Inspired Optimization Algorithm*s, Springer: Singapore, 2019, pp. 897–907. doi: 10.1007/978-981-13-0761-4_85.

[2] H. Aldowah, H. Al-Samarraie and W.M. Fauzy, "Educational data mining and learning analytics for 21st century higher education: A review and synthesis", *Telemat. Inform.*, vol. 37, pp. 13-49, 2019. doi: 10.1016/j.tele.2019.01.007.

[3] Ihor Tereikovskyi, Zhengbing Hu, Denys Chernyshev, Liudmyla Tereikovska, Oleksandr Korystin, Oleh Tereikovskyi, "The Method of Semantic Image Segmentation Using Neural Networks", *International Journal of Image, Graphics and Signal Processing*, vol. 14, no. 6, pp. 1-14, 2022.

[4] S.V. Balovsyak, O.V. Derevyanchuk, I.M. Fodchuk, "Method of calculation of averaged digital image profiles by envelopes as the conic sections", *Advances in Intelligent Systems and Computing*, Hu Z., Petoukhov S., Dychka I., He M. (Eds.), Springer International Publishing, vol. 754, pp. 204–212, 2019. doi: 10.1007/978-3-319-91008-6_21.

[5] S.V. Balovsyak, O.V. Derevyanchuk, I.M. Fodchuk, O.P. Kroitor, Kh.S. Odaiska, O.O. Pshenychnyi, A. Kotyra, A. Abisheva, "Adaptive oriented filtration of digital images in the spatial domain", *Proc. SPIE 11176, Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments*, vol. 11176, pp 111761A-1–111761A-6, 2019. doi: https://doi.org/10.1117/12.2537165.

[6] S.V. Balovsyak, O.V. Derevyanchuk, H.O. Kravchenko, O.P. Kroitor, V.V. Tomash, "Computer system for increasing the local contrast of railway transport images", *Proc. SPIE, Fifteenth International Conference on Correlation Optics*, vol. 12126, pp. 121261E1-7, 2021. doi: 10.1117/12.2615761.

[7] Derevyanchuk O.V., Kravchenko H.O., Derevianchuk Y.V., Tomash V.V. "Recognition images of broken window glass", *Proceedings of SPIE*", vol. 12938, pp. 210-213, 2024. doi: https://doi.org/10.1117/12.3012995.

[8] O. Berezsky, M. Zarichnyi, and O. Pitsun, "Development of a metric and the methods for quantitative estimation of the segmentation of biomedical images", *Eastern-European Journal of Enterprise Technologies*, vol. 6, no. 4(90), pp. 4–11, 2017. doi: 10.15587/1729-4061.2017.119493.

[9] Moez Ali. Clustering in Machine Learning: 5 Essential Clustering Algorithms. https://www.datacamp.com/blog/clustering-in-machine-learning-5-essential-clustering-algorithms.

[10] S.V. Balovsyak, Kh. S. Odaiska, "Automatic Determination of the Gaussian Noise Level on Digital Images by High-Pass Filtering for Regions of Interest", *Cybernetics and Systems Analysis*, vol. 54, no. 4, pp. 662-670, 2018. doi: 10.1007/s10559-018-0067-3.

[11] M. Wang, X. Wei, "Research on Logistics Center Location-Allocation Problem Based on Two-Stage K-Means Algorithms", *Advances in Computer Science for Engineering and Education*, Hu, Z., Petoukhov, S., Dychka, I., He, M. (Eds.), Springer International Publishing, vol. 1247, pp. 52-62, 2021. doi: 10.1007/978-3-030-55506-1_5.

[12] YanPing Zhao, and XiaoLai Zhou, "K-means Clustering Algorithm and Its Improvement Research", *Journal of Physics: Conference Series*, 2nd International Workshop on Electronic communication and Artificial Intelligence (IWECAI 2021) 12-14 March 2021, Nanjing, China, vol. 1873, 012074, pp. 1-5, 2021, doi: 10.1088/1742-6596/1873/1/012074.

[13] Guojun Gan, Chaoqun Ma, and Jianhong Wu, *Data Clustering: Theory, Algorithms, and Applications*, Second Edition. Society for Industrial and Applied Mathematics (SIAM), 2021.

[14] Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman, *Mining of Massive Datasets*, 3rd edition. Stanford University, 2019.

[15] X. Li, Y. Zhang, H. Cheng, F. Zhou and B. Yin, "An Unsupervised Ensemble Clustering Approach for the Analysis of Student Behavioral Patterns", *IEEE Access*, vol. 9, pp. 7076-7091, 2021, doi: 10.1109/ACCESS.2021.3049157.

[16] C. Maithri, H. Chandramouli, "Parallel DBSCAN Clustering Algorithm Using Hadoop Map-reduce Framework for Spatial Data", *International Journal of Information Technology and Computer Science*, vol. 14, no. 6, pp. 1-12, 2022.

[17] Mohamed Nafuri, Ahmad Fikri, Nor Samsiah Sani, Nur Fatin Aqilah Zainudin, Abdul Hadi Abd Rahman, and Mohd Aliff. "Clustering Analysis for Classifying Student Academic Performance in Higher Education", *Applied Sciences*, vol. 12, no. 19, pp. 9467, 2022. doi: 10.3390/app12199467.

[18] Denon Arthur Richmond Gono, Bi Tra Goore, Yves Tiecoura, and Kouame Abel Assielou, "Multi-relational Matrix Factorization Approach for Educational Items Clustering," *International Journal of Information and Education Technology*, vol. 13, no. 1, pp. 42-47, 2023. doi: 10.18178/ijiet.2023.13.1.1778.

[19] Mary Jane Samonte, Gabriel Edrick O. Acuna, Luis Antonio Z. Alvarez, and Jeffrey M. Miraflores, "A Personality-Based Virtual Tutor for Adaptive Online Learning System," *International Journal of Information and Education Technology*, vol. 13, no. 6, pp. 899-905, 2023. doi: 10.18178/ijiet.2023.13.6.1885.

[20] Zareen Alamgir, Habiba Akram, Saira Karim, Aamir Wali, "Enhancing Student Performance Prediction via Educational Data Mining on Academic data", Informatics in Education. vol. 23, no. 1, pp. 1-24, 2024. doi: 10.15388/infedu.2024.04.

[21] N. Shakhovska, O. Vovk, R. Hasko, Y. Kryvenchuk, "The Method of Big Data Processing for Distance Educational System", *Advances in Intelligent Systems and Computing*, Shakhovska N., Stepashko V. (Eds.), Springer International Publishing, vol. 689, pp. 461-473, 2018. doi: 10.1007/978-3-319-70581-1_33.

[22] Batool, S., Rashid, J., Nisar, M.W. et al., "Educational data mining to predict students' academic performance: A survey study", *Educ. Inf. Technol*, vol. 28, pp. 905-971, 2023. doi: 10.1007/s10639-022-11152-y.

[23] Sarsa, S., Leinonen, J., & Hellas, A., "Empirical Evaluation of Deep Learning Models for Knowledge Tracing: Of Hyperparameters and Metrics on Performance and Replicability", *Journal of Educational Data Mining*, vol. 14, no. 2, 2022. doi: 10.5281/zenodo.7086179.

[24] Nayak, P., Vaheed, S., Gupta, S. et al., "Predicting students' academic performance by mining the educational data through machine learning-based classification model", *Educ. Inf. Technol*, vol. 28, pp. 14611-14637, 2023. doi: 10.1007/s10639-023-11706-8.

[25] Delianidi, M., & Diamantaras, K. "KT-Bi-GRU: Student Performance Prediction with a Bi-Directional Recurrent Knowledge Tracing Neural Network", *Journal of Educational Data Mining*, vol. 15, no. 2, pp. 1-21. 2023. doi: 10.5281/zenodo.7808087.
[26] A. R. Fayek, "Fuzzy Logic and Fuzzy Hybrid Techniques for Construction Engineering and Management", *Journal of Construction Engineering and Management*, vol. 146, no. 7, pp. 1-12, 2020. doi: 10.1061/(ASCE)CO.1943-7862.0001854.
[27] Tengku Zatul Hidayah Tengku Petra, Mohd Juzaiddin Ab Aziz, "Analysing Student Performance In Higher Education Using Fuzzy Logic Evaluation", *International journal of scientific & technology research*, vol. 10, no. 01, pp. 322-327, 2021.
[28] A. Heni, I. Jdey and H. Ltifi, "K-means and fuzzy c-means fusion for object clustering", 8th International Conference on Control, Decision and Information Technologies (CoDIT), Istanbul, Turkey, pp. 177-182, 2022. doi: 10.1109/CoDIT55151.2022.9804078.
[29] Serhiy Balovsyak, Oleksandr Derevyanchuk, Vasyl Kovalchuk, Hanna Kravchenko, Maryna Kozhokar, "Face Mask Recognition by the Viola-Jones Method Using Fuzzy Logic", *International Journal of Image, Graphics and Signal Processing*, vol.16, no.3, pp. 39-51, 2024.
[30] Thamasan Suwanroj, Orawan Saeung, Punnee Leekitchwatana, and Kanaporn Kaewkamjan, "The Development of Professional Competency Test in Knowledge and Cognitive Skill for Computer Innovation and Digital Industry," *International Journal of Information and Education Technology*, vol. 13, no. 1, pp. 121-130, 2023. doi: 10.18178/ijiet.2023.13.1.1787.
[31] Omar T., Alzahrani A., Andzohdy M. "Clustering approach for analyzing the student's efficiency and performance based on data", *Journal of Data Analysis and Information Processing*, vol. 8, no. 03, pp. 171–182, 2020. doi: 10.4236/jdaip.2020.83010.
[32] J.C. Russ, *The Image Processing Handbook*. Taylor and Francis Group, 2011.
[33] Google Colab. URL: https://colab.research.google.com.
[34] Derevyanchuk O.V., Balovsyak S.V. Certificate of copyright registration for the work, No. 123369, 31.01.2024. ID CR3206310124. Computer program " Data clustering using fuzzy logic" ("ClusterFuzzy23"). Ukrainian National Office for Intellectual Property and Innovations (IP Office). URL: https://sis.nipo.gov.ua/uk/services/original-document.
[35] scikit-learn. URL: https://scikit-learn.org.
[36] Students performance in exams. URL: https://www.kaggle.com/datasets/spscientist/students-performance-in-exams.
[37] M. Zhang, Y. Cheng, J. Gu, Y. Yang, L. Chen, and B. Cui, "Digital Evaluation of Innovative Logistics Talents Based on Improved SAGA-FCM Algorithm," In: Hu, Z., Zhang, Q., He, M. (eds), Advances in Artificial Systems for Logistics Engineering, Springer, Cham., vol. 223, pp. 374-383, 2024. doi: 10.1007/978-3-031-72017-8_34.
[38] Nadindra Dwi Ariyanta, and Anik Nur Handayani, "Sugeno Fuzzy Personality Prediction System: An Approach to Overcoming Psychological Measurement Uncertainty," *Indonesian Journal of Data and Science*, vol. 5(3), pp. 216-228, 2024. doi: 10.56705/ijodas.v5i3.192.
[39] Free Personality Test. URL: https://personalitymax.com.
[40] Experience Based Learning Systems, LLC (EBLS). URL: https://learningfromexperience.com.
[41] Kolb's Learning Styles and Test. URL: https://en.eduolog.com/kolbs-learning-styles.

Authors' Profiles

Zhengbing Hu: Prof., Deputy Director, International Center of Informatics and Computer Science, Faculty of Applied Mathematics, National Technical University of Ukraine "Kyiv Polytechnic Institute", Ukraine. Adjunct Professor, School of Computer Science, Hubei University of Technology, China. Visiting Prof., DSc Candidate in National Aviation University (Ukraine) from 2019. Major research interests: Computer Science and Technology Applications, Artificial Intelligence, Network Security, Communications, Data Processing, Cloud Computing, Education Technology.

Oleksandr Derevyanchuk: Received the Master of Engineering degree (1999) and Ph.D. of Physics and Mathematics (2014) at the Yuriy Fedkovych Chernivtsi National University. He is a Candidate of Physical and Mathematical Sciences, Associate Professor of the Department of Professional and Technological Education and General Physics, Yuriy Fedkovych Chernivtsi National University, Chernivtsi, Ukraine.

Research Interests: Education Technology, Educational Data Mining, Digital Processing of Signals and Images, Programming, Pattern Recognition, Artificial Neural Networks.

Serhiy Balovsyak: Graduated from Chernivtsi State University (1995). In 2018, he defended his doctoral dissertation in the specialty "Computer systems and components".

Currently position – associate professor at the Department of Computer Systems and Networks of Yuriy Fedkovych Chernivtsi National University, Ukraine.

Research Interests: Education Technology, Data Mining, Digital Processing of Signals and Images, Programming, Pattern Recognition, Artificial Neural Networks.

Yuriy Ushenko: M.Sc. in Telecommunications (2003). PhD in Optics and Laser Physics (2006). D.Sc. in Optics and Laser Physics, Taras Shevchenko National University of Kyiv (2015).

Current position – Professor, Head of Computer Science Department, Yuriy Fedkovych Chernivtsi National University, Ukraine.

Research Interests: Data Mining and Analysis, Computer Vision and Pattern Recognition, Optics & Photonics, Biophysics.

Hanna Kravchenko: Teacher of the State Educational Establishment "Chernivtsi transport college", Chernivtsi, Ukraine.

Research Interests: Education Technology, Educational Data Mining, Digital Processing of Signals and Images, Programming, Pattern Recognition, Artificial Neural Networks.

Iryna Sapsai: M.Sc. of Physics M. Dragomanov National Pedagogical University of Kyiv (2009). PhD in Pedagogic "Physics", M. Dragomanov National Pedagogical University of Kyiv (2014). Current position – Lecturer at the Department of Science and Mathematics Education and Technology, Institute of Postgraduate Education, Borys Grinchenko Kyiv Metropolitan University.

Research Interests: Physics education, Optics, Teaching and Learning, Pedagogy and Education, E-learning.

How to cite this paper: Zhengbing Hu, Oleksandr Derevyanchuk, Serhiy Balovsyak, Yuriy Ushenko, Hanna Kravchenko, Iryna Sapsai, "Fuzzy Clustering of Educational Data with Automated Selection of Processing Parameters in System Analysis of Quality Education", International Journal of Modern Education and Computer Science(IJMECS), Vol.17, No.4, pp. 101-111, 2025. DOI:10.5815/ijmecs.2025.04.07