Intelligent Data Clustering System for Searching Hidden Regularities in Financial Transactions

Nickolay Rudnichenko[a], Vladimir Vychuzhanin[a], Natalia Shibaeva[a], Igor Petrov[b], Tetiana Otradskya[a]

[a] *Odessa Polytechnic National University, Shevchenko Avenue 1, Odessa, 65001, Ukraine*
[b] *National University "Odessa Maritime Academy", Didrichson street 8, Odessa, 65029, Ukraine*

Abstract

The article presents results of the intelligent data clustering system for searching hidden regularities in financial transactions development. The main aspects and problems of increasing the volume of financial information within the client base segmentation for the formation of various development strategies and marketing methods development for promoting goods in order to expand the target audience are given. The key opportunities and difficulties of using modern data mining methods and algorithms based on supervised and unsupervised learning are described and analyzed. Existing hybridization approaches implementation for data analysis algorithms, including those based on the use of data clustering ensembles, are considered. The concept of data analysis stages in the process of solving the segmentation problem is proposed, research metrics are formalized, clustering algorithms are selected and programmatically implemented via information system with the assignment clusters initial number and calculating it independently. Collected and formed balanced set of data on financial transactions for research, performed its statistical analysis, transformation and preparation for clustering. A software implementation of the system has been developed and its key functionality, component composition has been designated. The developed algorithms results studies based on the summary matrix of feature proximity analysis are presented, a unified space for cluster visualization is created based on the t-SNE approach, clustering quality assessing metrics are calculated.

Keywords

Cluster data analysis, hidden patterns search, segmentation, financial transactions, data mining.

1. Introduction

Currently, there is a rapid increase in the number of non-cash financial transactions, both using payment cards and through global payment systems with private currency units, electronic money, and cryptocurrencies. This is especially noticeable in the global consumer electronic commerce market through the use of online stores, social networks and other channels for the sale of products, the provision of services or transfers between individuals or legal entities, in connection with which the cash form of payment for goods and services is becoming less and less popular and up-to-date [1]. This trend is largely due to the convenience for the end user of the processes of conducting payment transactions through the use of mobile applications or Internet services for targeted payment, electronic billing systems, well-coordinated work of issuer banks, prompt and secure forms of acquiring transactions.

Making payment non-cash transactions process within the software or information systems used, heterogeneous data sets are formed that represent the history of purchase transactions with reference to time, location, type or costs category [2].

ICST-2023: Information Control Systems & Technologies, September 21-23, 2023, Odesa, Ukraine.
EMAIL: nickolay.rud@gmail.com (N.Rudnichenko); vint532@gmail.com (V.Vychuzhanin); nati.sh@gmail.com (N. Shibaeva) firmn@gmail.com (I. Petrov); tv_61@ukr.net (T. Otradskya).
ORCID: 0000-0002-7343-8076 (N.Rudnichenko); 0000-0002-6302-1832 (V.Vychuzhanin); 0000-0002-7869-9953 (N. Shibaeva); 0000-0002-8740-6198 (I. Petrov); 0000-0002-5808-5647 (T. Otradskya).

Such data can be valuable for businesses in terms of intelligent analysis (DM) of hidden patterns based on the use of machine learning (ML) deep learning (DL) to build predictive models or customer segmentation systems (by a number of characteristics, for example, by target audience or digital footprint, portrait or behavioral image) in certain areas of human activity [3].

Identification of hidden patterns in the nature of payment transactions allows us to simplify and improve the marketing process. In particular, the segmentation of the client base by separate groups is advisable for the formation of personalized and most relevant to user expectations of promotional offers, discount programs, special conditions and accumulative bonus systems, providing a higher level of services provided as part of the competition for customers between players in the market of goods and services [4].

The scientific and applied side of existing research in the designated area can be useful in the context of development: directly of ML or DL methods and algorithms; development of analytical systems and working environment capable of optimizing computing processes for data processing; visualization and analysis results interpretation [5].

The complexity transactional data analysis lies in a number of aspects related both to the policy of organizations that collect such information on its formation and provision, and to the technical imperfection of the systems and technologies used. In particular:

1. in order to protect confidential information, some of the personal data of users describing the completed transaction, which clearly identify the person, are hidden or inaccessible even at the stage of their receipt, which complicates the formation of an exhaustive list of significant input features [6];
2. the volume of generated data strongly depends on the data collection system, the number of active clients, the capabilities and relational or non-relational storages settings used;
3. in large companies, the volume of collected data is dynamically growing, which complicates the procedures for storing and maintaining data online, and the process of manual or automated data labeling in some cases is laborious or inappropriate, which does not allow the effective use of a number of ML supervised algorithms [7];
4. in a number of systems there is a problem of poor data ordering, which entails the periodic formation of missing values, incorrect transactions due to failures in banking or third-party payment systems, terminals, which complicates the data processing process and requires additional procedures for their cleaning and preparation to the analysis [8];
5. there is a need to implement additional resource-intensive procedures for filtering and sorting data in order to identify and remove anomalous values or outliers that may occur if customers or payment service support operators enter erroneous or incorrect data during transactions;
6. there are no effective and proven in practice in such tasks intelligent systems, methods and algorithms that provide flexibility in the interpretation of the results obtained, as well as fully taking into account the specifics and nature of financial transactions, capable of providing operational data analysis with sufficient levels of accuracy and speed [9].

All this determines the relevance of studying the possibilities of using hybrid intelligent ML technologies, methods and systems for analyzing financial transactions in order to search for hidden patterns in targeting and segmenting the customer base.

2. Description of Problem

An analysis of modern publications focused number on DM questions on customer segmentation of various organizations according to a input features number, including in the financial sector, made it possible to establish the fact that approaches in this area are predominant based on the use of algorithms and supervised or unsupervised models [7, 9].

In the first case, the authors [10] use classical ML algorithms, but their complexity is the flexibility in interpreting the results in the presence of a target features large number, which does not allow taking into account the specific aspects of the data obtained during their collection and processing. This, in turn, requires an additional solution to the problem of data dimensionality reduction, which is quite laborious and not always efficient based on the use of existing functional applied libraries in the field of data analysis [11]. Some authors use rather complex and cumbersome

DL models built on the basis of a combination of multilayer neural networks, which gives a number of positive results, but the models do not have sufficient generalizing ability to correlate data with target audience groups to build practically useful forecasts and make managers. solutions [12].

Most often, the authors in practice reduce the problem to considering the specifics of the multiclass classification problem [10,13,14], in which the data are initially marked by an expert or based on automated similarity markup tools. These approaches have several disadvantages, in particular:

1. the use of ML methods separately does not allow achieving sufficiently high results;
2. significant time spent on preparing and processing data due to the need for manual verification, validation and labeling of all records;
3. those formed within the framework of the model are able to predict the values of the output variable for individual classes, but they do not allow revealing hidden patterns in the data and taking into account non-obvious signs;
4. markup automation is not always performed correctly and requires the configuration of individual tools, modules and dependencies for a specific task, which is not always justified in terms of computing resources.

It should also be noted the different efficiency of the approaches used for different data sets, which largely depends on the types of algorithms and metrics used to verify and evaluate ML models quality [15].

In the case of using ML algorithms without a teacher, a number of identified problems are leveled, but new aspects of model formation arise that need to be addressed. In particular, a number of classical algorithms, for example, k-means, are fast in terms of computational performance, but most often they do not allow achieving a sufficiently high accuracy level [8,11,14]. In this regard, their additional modification or optimization is required to ensure the consistency of the results.

The mathematical formalization of the clustering problem can be represented as follows. Let X be a set of objects and Y be a set of numbers or labels of individual clusters. The given function for determining the distance between the available objects is a training sample of objects $X^m = \{x_1,...,x_m\} \subset X$.

It is required to split the existing sample into non-overlapping subsets, i.e. clusters, so that a separate cluster consists of various objects that are close in metric p, and objects belonging to different clusters differed significantly. For each object $x_i \in X^m$ a separate cluster number is assigned y_i.

The clustering algorithm is the function $a: X \to Y$, which for each object $x \in X$ puts in direct correspondence the number of a separate cluster $y \in Y$. Set Y in some cases it is known in advance, but it is necessary to determine the optimal number of clusters, in terms of a given criterion for the quality of clustering results [4,5].

It should be noted that, from our point of view, the most appropriate approach is the analyzing data hybridization process on financial transactions based on a various combinations clustering algorithm [16]. Those, within the framework of the problem under consideration, it is advisable to use unsupervised learning algorithms, in particular methods of cluster data analysis, the advantages of which for the analysis of financial transactions are:

- no need to markup financial data and simplify their pre-processing;
- the ability to combine several algorithms within an ensemble in order to obtain better results and metrics;
- the possibility of combining clustering algorithms, in which the values of the number of output clusters are initially set to select the desired target features in different ranges, as well as algorithms with automatic calculation of the number of clusters, which allows to ger results with additional verification;
- eliminating the problem of introducing inaccuracies into the analysis results due to the presence of outliers;
- distribution of computations among individual clusters based on the use of multi-threaded approaches when conducting experiments.

An additional advantage of this approach is the achievement of greater stability of the resulting solutions based on the adjacent ensembles algorithms formation by constructing a collective solution with many opinions (options for splitting data into clusters). The application of this approach makes it possible to reduce the dependence of the grouping results on the algorithm parameters choice, to develop more stable solutions in noisy data. The collective decision function combines each methods advantage used in the function's construction [17].

When developing a collective decision, the results obtained by each individual model are grouped, because the more successful solution compensates for the less successful ones. At the same time, stable patterns, in accordance with which clusters are formed, are mutually enhanced, while unstable ones are weakened. The indicated approach makes it possible to carry out distributed computing in the case when obtaining a common database is impossible or not economically feasible. These properties of cluster analysis are especially relevant when working in areas with intensive data use, which is typical for the financial market [18].

Summarizing the results obtained, it is necessary to single out the main stages of cluster data analysis during the segmentation operation (in general, they are shown schematically in Fig. 1): selecting a sample for data clustering; determination of the required set of variables by which the objects in the data sample will be evaluated; calculation of values of the degree of similarity between the analyzed objects; application of a specific clustering approach to form groups of similar objects; results quality control.

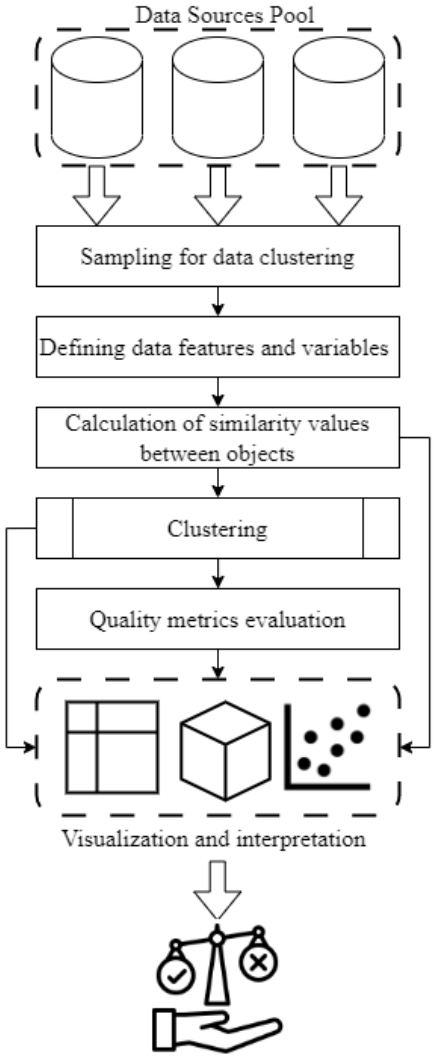

Figure 1: Data analysis stages in segmentation

An additional relevant argument in favor of using this approach and developing our own system is ready-made data analysis systems lack that allow us quickly and accurately ensure clustering

formation process with necessary algorithms selection. The currently existing clustering software systems in most cases are either private research works, which usually have fixed (not expandable) sets of algorithms, or commercial products, which are primarily focused on corporate clients use, which makes it impossible to have them for research and analysis [19].

Thus, it should be noted that the purpose of the work, which is to develop and study an intelligent system for identifying hidden patterns in transactional financial data, is relevant and in demand in practice.

3. System concept and development

The developed system is aimed at revealing hidden patterns in financial transaction data using the customers RFM segmentation by different types of groups in order to formalize the forming target audience process for further promotional offers development, demand management and customer base retention.

3.1. Ensemble of clustering algorithms

Based on studies comparing a number of clustering algorithms popular in practice [17-23], it was decided to use 4 algorithms within the system being designed, 2 of which require an explicit specification of the number of clusters, and the remaining 2 calculate this value automatically.

The k-means algorithm, the idea of which is to minimize the distances between corresponding objects in separate clusters. The calculation process is stopped under the condition when minimization is impossible. Minimization by the k-means algorithm is carried out based on the following expression

$$J = \sum_{k=1}^{M} \sum_{i=1}^{N} d^2(x_i, c_k), \tag{1}$$

where $x_i \in X$ - clustering object (single point in space);

$c_k \in C$ - selected cluster center.

FOREL is a clustering algorithm based on the idea of combining objects into one cluster in the areas of their greatest concentration. The algorithm purpose is to divide the sample into an initially unknown clusters value so that the sum of distances from cluster objects to their centers is minimal. In fact, as algorithm implementation part it is necessary to select groups of objects as close as possible to each other, which, due to the similarity hypothesis, will form the final clusters.

$$F = \sum_{j=1}^{k} \sum_{x \in K}^{N} p(x, W_j) \tag{2}$$

The first summation is carried out over all sample clusters, the second summation is over all objects x, belonging to the current cluster K_j, where W_j - current cluster modified center, $p(x, y)$ - distance between objects. DBSCAN is a density clustering algorithm, defined as a radius and the number of sample points in the generated range. In DBSCAN, each point (data object) in the sample is assigned one of the labels: the central point or the boundary (according to the specified radius R). Based on each core point or group of related core points, a separate cluster is formed (in the case when the distance between the core points is very close), after which each of the boundary points is divided into corresponding core points groups. Ward's hierarchical algorithm combines the results of previous algorithms to form a sample final clustering dendrogram. The general concept of forming an algorithms ensemble (fig.2) is based on the bagging approach, i.e. parallel models' creation and their use on random data subsamples taken from a common pool. The final result definition for the objects in the identified clusters is determined by the voting of all ensemble models.

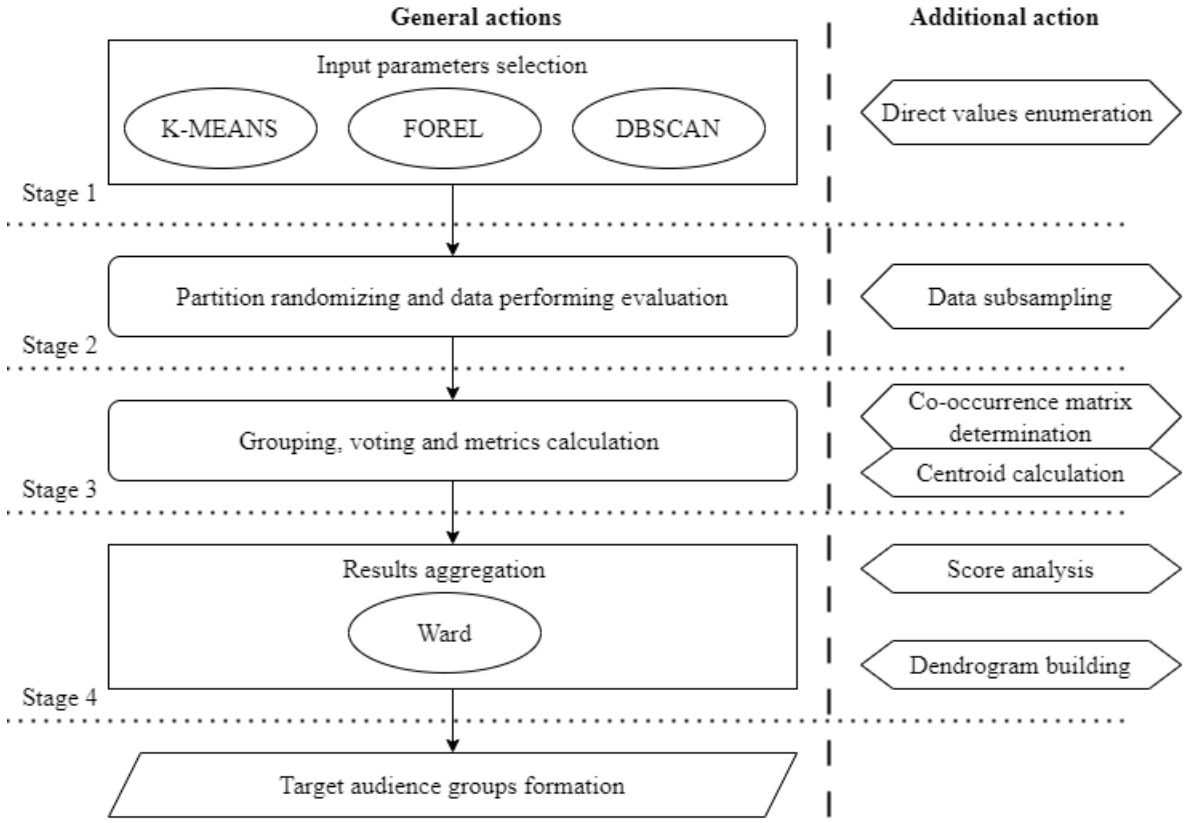

Figure 2: Formed clustering ensemble work scheme

The criterion for calculating the degree of similarity and clusters difference is the distance between the available points on the compiled scatter diagram, which corresponds to the distance between the points located on the graph. Within developed system framework, the method for estimating the Euclidean distance between points *i* and *j* on the plane is used (D_{ij}).

$$D_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{3}$$

3.2. Metrics evaluation

To assess the accuracy and reliability of the proposed approach to clustering, the following metrics were chosen. Silhouette coefficient (S). Unlike other metrics, this measure does not imply initial knowledge of the true labels of the analyzed objects, and therefore allows us to evaluate cluster partitioning quality on an unlabeled sample, forming the final clustering result. Initially S defined separately for each object in the imported data selection. If a is the average distance from the given object to other particular cluster objects, b is the average distance from a given object to objects located in another (nearest) cluster, then the silhouette of this object will be the following value

$$s = \frac{b - a}{\max(a, b)}. \tag{4}$$

In fact, the sample silhouette will be the objects silhouette average value included in a given data set. The convenience of this metric lies in the fact that it allows us to set the difference degree between the average distance to the cluster objects from the average distance to the objects of other clusters into which the sample is divided. The value of this metric is also in the range from -1 to 1, and the lower the score, the worse the clustering is performed (it is scattered). To determine the clusters optimal number Elbow score method is used, which is a heuristic used to determine the

number of clusters in a data set. This method considers the change in the spread W_{total} with an increase in groups number k. By combining all *n* observations into one group, we can get the largest intracluster variance, which will decrease to 0 when $k \to n$.

$$W_{total} = \min(k \in (1; n)) \tag{5}$$

Kalinsky-Harabas index (for algorithms without explicitly specifying the number of clusters, i.e. FOREL and DBSCAN, since this metric allows you to additionally assess the degree of applicability of these algorithms to the selected data set), which can be used to determine the optimal value for clustering. The Kalinske Harabas index is defined as

$$VRC_K = \frac{SS_B}{SS_W} \times \frac{(N-K)}{(K-1)}, \tag{6}$$

where K - clusters number, N - samples number, SS_B - squares error sum between groups, a SS_W - squares error intragroup sum. An additional factor is the algorithm profiling in the execution time estimation form.

3.3. Dataset description

As an input data set, information was collected on customers financial transactions in online commerce, detailing their behavioral functions and reactions based on the open source number usage, including UNBANKS Real-Time EU Consumer Transaction Data. When the problem of unsupervised learning is solved, then the data are not initially labeled, so it is necessary to evaluate features similarity (proximity) to each other. For this purpose, a factor analysis was initially carried out to identify the most significant features to identify hidden patterns, the number of features was reduced from 255 to 15. Input features structure is divided into groups for the clustering process convenience.

1. Customers (general customer data): customer_id - client ID; product_X - product status. OPN - opened but not disposed of. UTL - reclaimed. CLS - closed; gender_cd - gender of the client. M - male. F - female; age - client's age in years; marital_status_cd – client's family status (MAR - married, CIV - civil marriage, DLW - does not live with spouse, UNM - single / not married, DIV - divorced, WID - widower, widow); children_cnt - client children number; first_session_dttm - date and time of the first session in the application or personal account; job_position_cd - category of the position held (1 - service personnel; 2 - own official business; 3 - own unofficial business; 4 - head of the organization; 5 - senior manager; 6 - category not specified; 7 - housewife; 8 - individual entrepreneur; 9 - disabled person ; 10 - head of the initial level; 11 - temporarily unemployed; 12 - old-age pensioner; 13 - disability pensioner; 14 pensioner; 15 - seniority pensioner; 16 - non-managing specialist employee; 17 - head of the unit; 18 - unknown ; 19 - works for individual entrepreneurs; 20 - non-executive employee-worker; 21 - working pensioner); job_title - client's position.

2. Transactions (data on completed transactions): customer_id - client ID; transaction_month - month of transaction; transaction_day - transaction day; transaction_amt - transaction amount in dollars; merchant_id - ID of the store where the transaction was made; merchant_mcc - MCC category.

The total data sample size is about 50,000 customer records and more than 1 million transaction records. For pre-processing of the collected raw data, they are aggregated, recorded and saved in *.txt format, and also structured as two separate samples according to customer data and transactions for sequential transaction analysis convenience. After saving and formatting the raw data in the *.csv format, they can be visualized in a tabular form convenient for interpretation and analysis. In order to conduct a preliminary exploratory analysis, a statistical study of the collected data was carried out, fragments of the results for the first sample are shown in Fig. 3.

The prepared data is divided into equal fragments for the purpose of their distributed processing within ensemble using framework to simplify the procedure for their random distribution between clustering models.

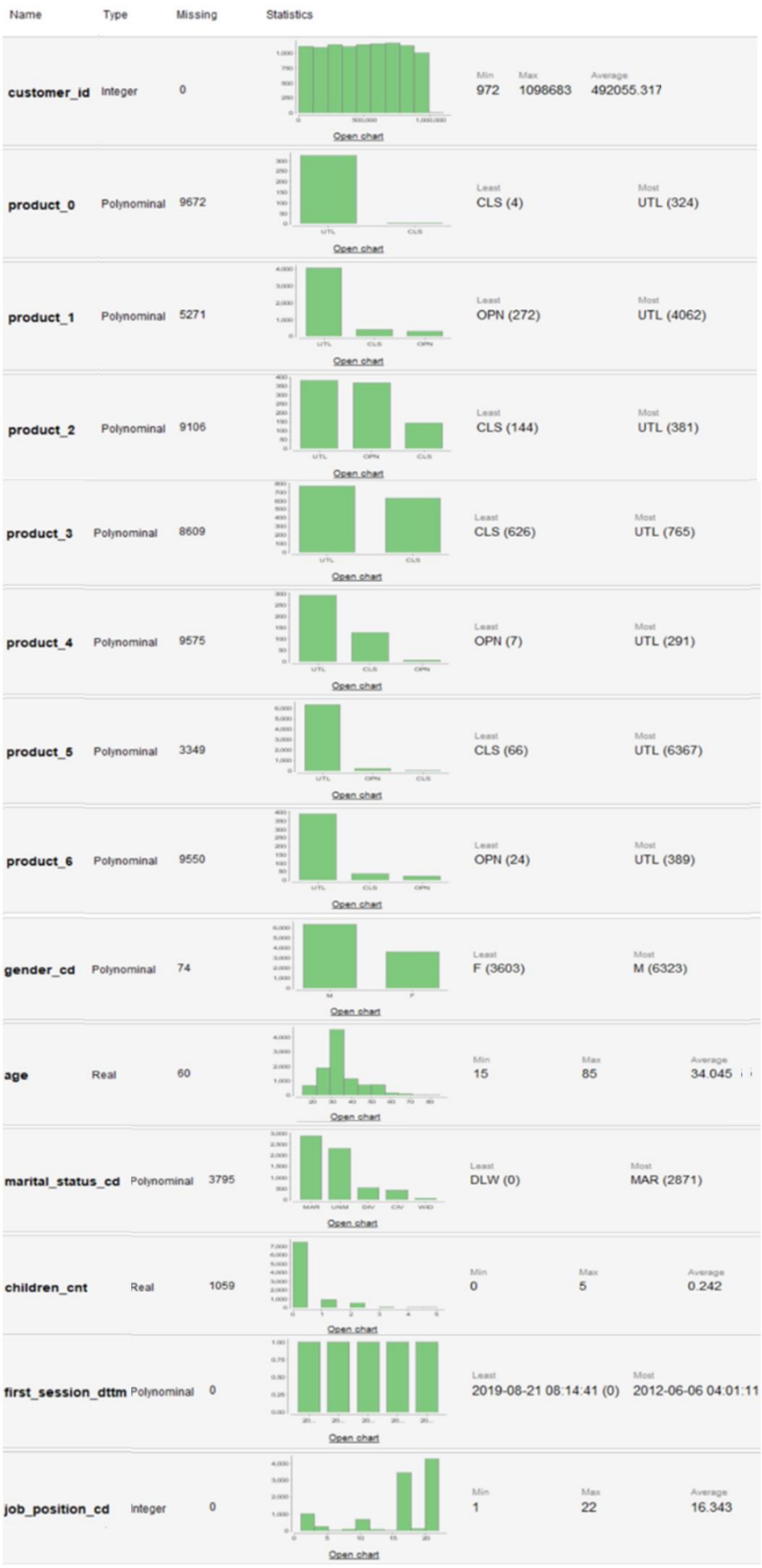

Figure 3: Summary analysis of sample data

3.4. Software development

For object-oriented software application operation modeling, we use UML methodology and notations. The main options developed diagram for software application using is shown in Fig. 4.

The user can carry out: import of input data sets; factor analysis; preprocessing and cleaning of data from outliers; input features choice; building a diagram by features; filling in gaps; formation of summary statistics; data normalization; data clustering by one or more of the selected algorithms (Forel, k-means, Ward, DBScan) as assemble; building visualizations (dendrograms, error matrices and scatter plots); assessment of clustering metrics; formation of a table of segmentation results.

After importing the data, the input datasets are loaded, collections are formed to store the imported records, after which exploratory analysis is performed, including feature analysis, the formation of additional average features for data comparison, and dataset key features distribution visual diagrams construction.

Based on the formed data structure, missing values in the data set are eliminated, records are aggregated into a single set. After that, the normalization objects are set for each of the analyzed features. Each clustering algorithms parameters are configured based on the performed normalization, after that the distances between the features in the generated state space are estimated. Next step is clustering metrics calculation and the results are displayed in graphical visualizations form.

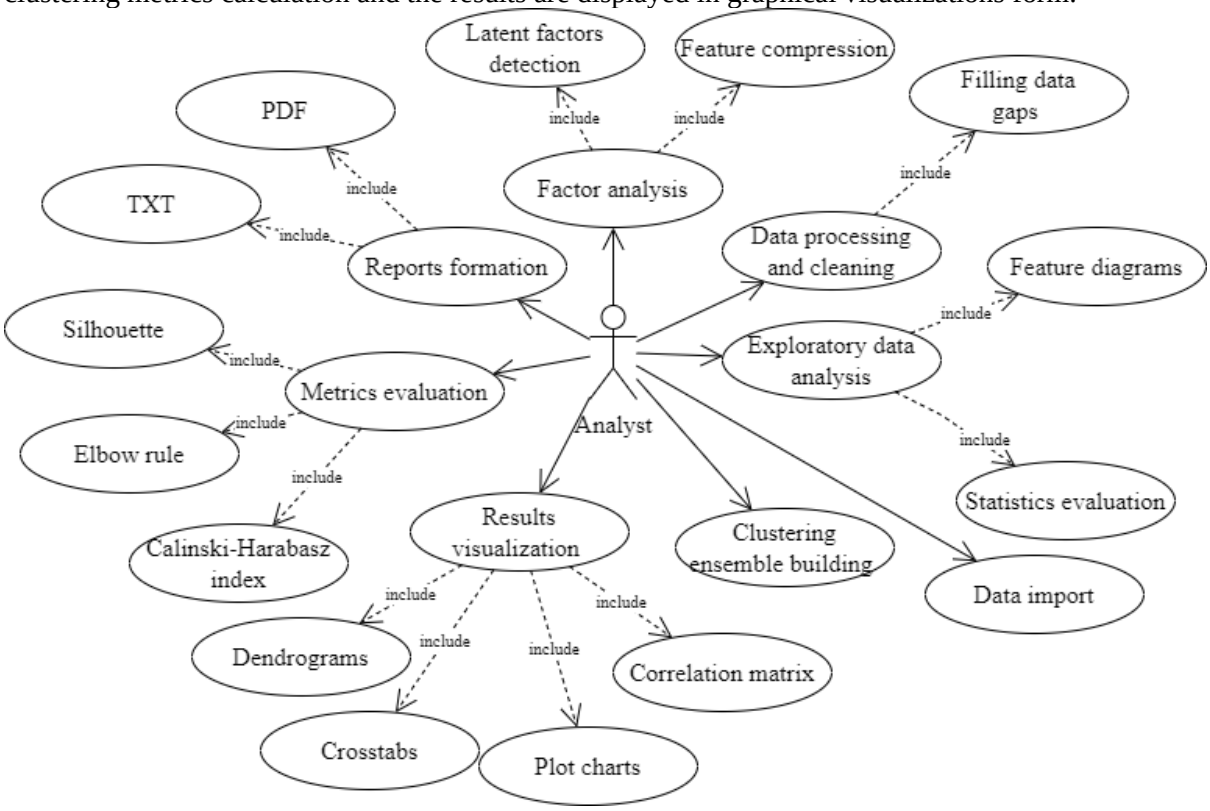

Figure 4: Use Case Diagram

The main application, when forming an clustering algorithms ensemble, uses a number of external dependencies (scipy for hierarchical clustering; sklearn, NumPy, Pandas for building data structures and clustering algorithms; matplotlib and seaborn for visualizing clustering results in a graphical form) and own FOREL and Transformers objects for implementation Forel algorithm logic and data normalization, respectively. The application deployment diagram is shown in Figure 5.

Local development is done using the Python programming language version 3.8 and the PyCharm development environment. By using the Google Chrome browser, when authorizing through a gmail account, one enters the Google Colab cloud web service for deployment, testing and research of the system in the form of an *.ipynb file. For convenience, data is loaded directly from the cloud storage directly into the system upon request.

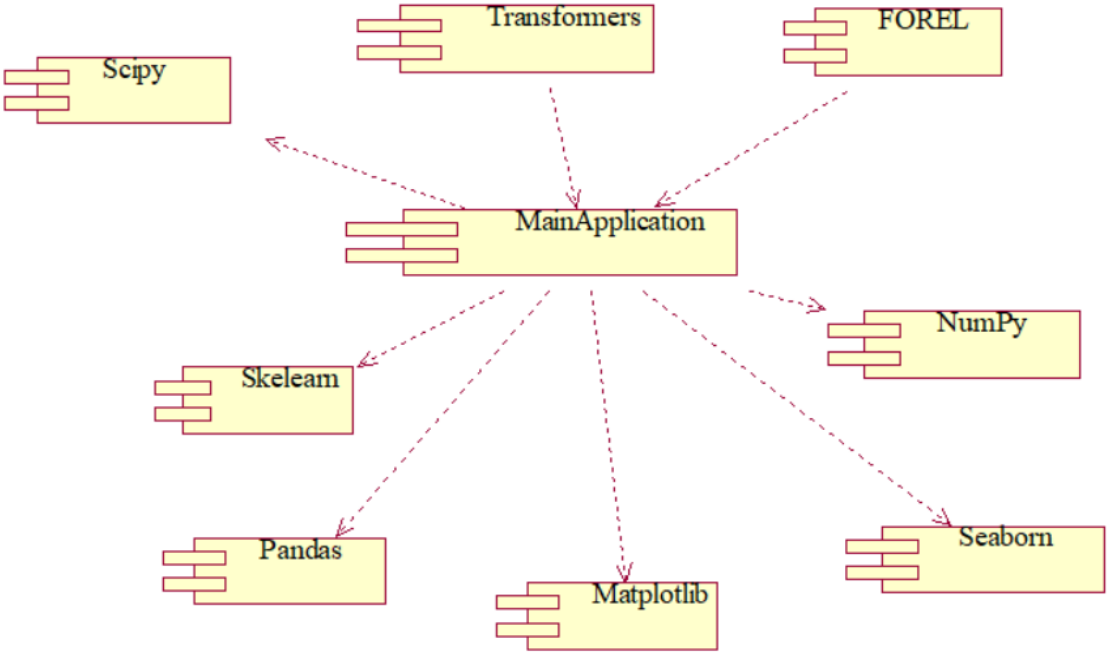

Figure 5: Main Components

Within the framework of the developed system, it imposes a number of requirements on input data samples, which is why a number of preliminary checks are implemented when importing data, in particular:

- verification of the absence of a direct correlation between features (correlation coefficient must be less than a user-specified threshold, 0.5 by default);
- verification of dimensionless form of data;
- distribution nature assessment for similarity to normal;
- indicators stability;
- sample homogeneity.

4. Experiments and results analysis

In order to carry out a created clustering ensemble work generalized assessment a number of studies were carried out on various transactions volumes. For the k-means algorithm, the values of the following hyperparameters are set: the number of clusters n_clusters in the range from 4 to 9, the number of setting different centroids n_init = 'auto', the computational iterations maximum number value max_iter in the range from 200 to 300, relative tolerance tol = 1e-5 , a variation of the algorithm chosen by lloyd and elkan.

The following hyperparameters are specified for the DBScan algorithm: the maximum distance between two samples eps in the range from 0.3 to 0.7; number of samples in the vicinity of the point min_samples = 4; proximity calculation metrics metric = euclidean; version of the comparison algorithm algorithm = brute; number of simultaneously running computing processes n_jobs = 4.

For the Forel algorithm, the values of the search radius parameter for local concentrations were set using sliding control.

On the basis of the conducted studies, input data individual categories proximity indicators relative to each other were formed and evaluated and are expressed in relative units.

The generalized matrix for average values analysis by features in clusters relative to the global average is shown in Fig.6. Clusters numbers selected by the algorithms are plotted along the y-axis, the main features are indicated along the x-axis, and a color scale is additionally introduced to visually display the features similarity degree in clusters.

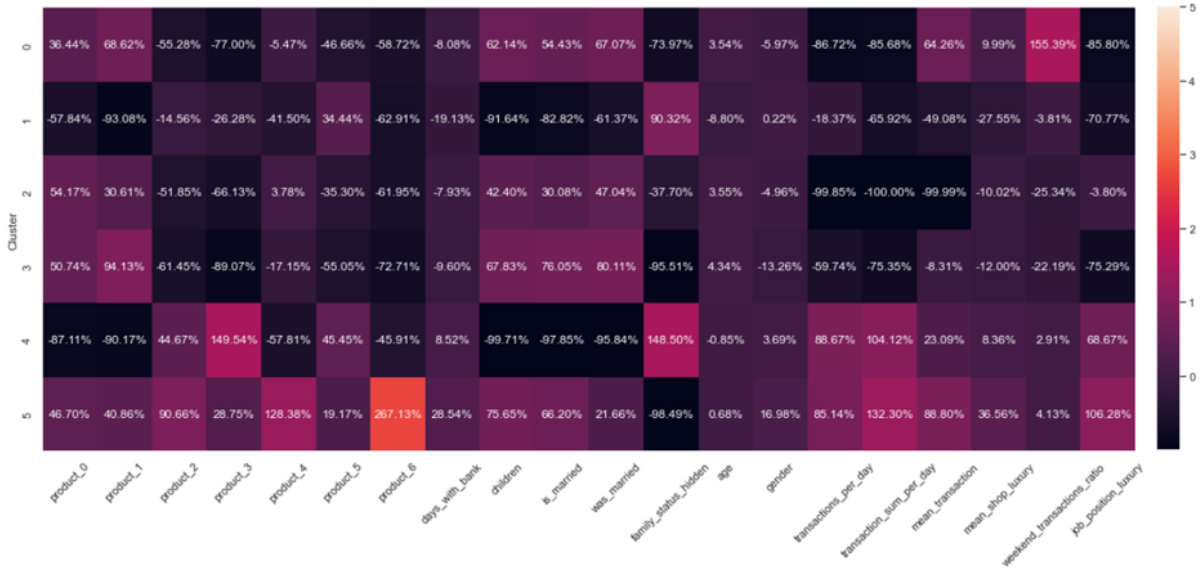

Figure 6: Generalized matrix for the mean values proximity by features in clusters relative to the global mean

Some fuzziness between 4 and 5 clusters should be noted. Due to the fact that the results of the clustering ensemble are not linearly separable data due to large dimensions presence, it is necessary to compose mixtures of objects that will belong to normal distributions with different parameters using the t-SNE algorithm, which will ensure multidimensional space decomposition to the appropriate level. This process is resource intensive and was performed in a distributed mode based on the use of 4 separate threads and took about 50% of the total duration of all computational experiments. Result visualization obtained based on developed clustering ensemble use is shown in Fig. 7.

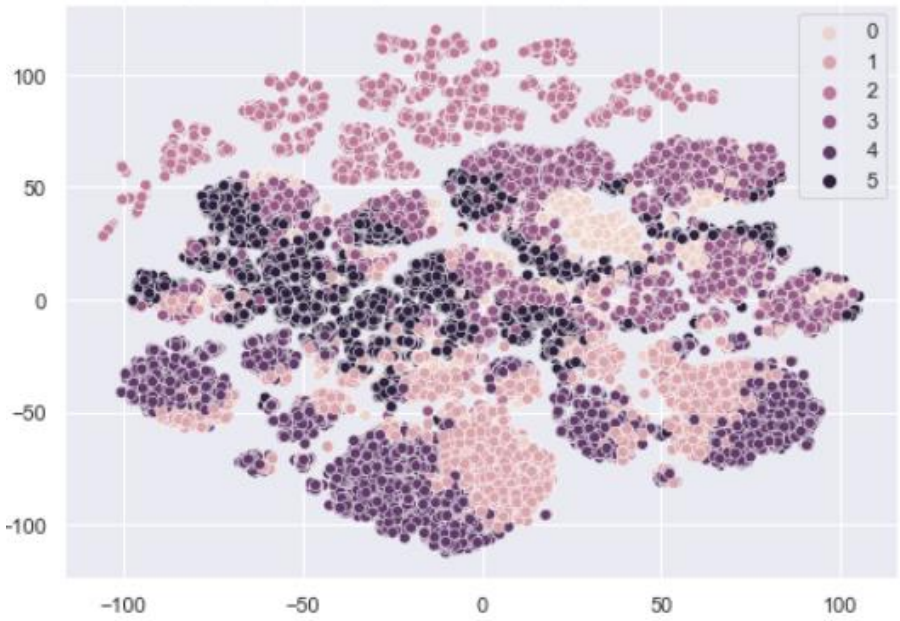

Figure 7: The result of the final space decomposition by dimension based on t-SNE

As we can see from the results obtained, clusters 2 and 1 are more focused and separated from the rest. The greatest blurring character corresponds to clusters 3 and 5, which may be due to the correlation between a number of secondary features that affect the generated transactions and are not separated into separate subgroups. Based on a summary assessment of all algorithms implemented within the clustering ensemble, a table was formed with the obtained metric values results assessments and overall performance in the computing operations profiling time form (Table 1).

Table 1
Results of metrics calculation

	K-means			DBScan			FOREL		
Cluster / Group	Clients number	S	Time-cost (sec)	Clients number	S	Time-cost (sec)	Clients number	S	Time-cost (sec)
0	13069	0,87	254	12056	0,92	561	12467	0,88	333
1	30667	0,78	412	29887	0,83	1359	30959	0,86	877
2	5960	0,79	114	7770	0,86	322	6281	0,8	163
3	149	0,64	2	165	0,77	4	156	0,83	2
4	134	0,81	2	111	0,79	2	124	0,9	2
5	21	0,9	1	11	0,91	1	13	0,81	1

Due to the specifics of the analyzed data and taking into account a hidden dependencies number, clustering algorithms do not always clearly and unambiguously segment customers. However, the analysis of the operation of clustering ensemble algorithms made it possible to identify some commonality in a number of features, which is the basis for identifying a customer's categories number. In particular, categories can be distinguished by 6 user clusters (presence/absence of children, over/under 45 years old, exceeding/not exceeding the average monthly salary for the position, male/female, married/never married, shopping on weekdays/weekends).

Based on the results obtained within the considered data set, it is reasonable to note that the identified 6 customer clusters, taking into account their purchases according to the MCC code, should be attributed to the following named categories: housewives, businessmen, models, transport employees (logistics), tourists and students (students) .

Within the framework of the developed system, it is easier and more convenient to configure and evaluate the k-means clustering algorithm (which is the fastest, but there is a greater spread in the values of metric estimates), since dbscan and FOREL are highly sensitive to changes in hyperparameters, and therefore their time costs are higher. The k-means and FOREL algorithms provide more balanced clusters. Analyzing the differences in average values in different clusters, we can see that the algorithms often separate customers by product (for example, product 6), but not so often by gender. Customers are better separated by transaction-related parameters, such as how expensive the goods are in stores, and how much money the customer spends relative to other customers in a similar role. By visualizing average values in clusters relative to global averages, it is possible to identify relationships between the several parameters manifestations at once, especially if these relationships are detected by several algorithms at once.

The relative maximum discrepancy between the algorithms in clients number terms in the cluster is quite insignificant and is about 3.5% for group 2, the smallest discrepancy values are noted for groups 1 and 5. Such values are an additional side sign input features choice correctness as a result of factor analysis and as a clustering high accuracy result.

5. Conclusion

The approach to data analysis proposed in this article and the developed system that provides software support for the described algorithms with the ability to customize them and build an clustering assemble, makes it possible to automate identifying hidden patterns process in data by combining unsupervised learning algorithms. The use of hierarchical and non-hierarchical data clustering based on the ML models' ensembles usage allows multi-level customer segmentation in a multidimensional feature space.

The created system can be used for data segmentation tasks with different structure and application area due to its versatility in data preprocessing functionality terms and results visualization.

A further target proposed approach developing way is the integration within the system of other clustering data analysis algorithms, as well as methods for forming clustering ensembles based on

boosting or stacking, followed by the development of a functional to compare their effectiveness with each other on the same data sets. Also, it is advisable to supplement the system with the ability to support other metrics for evaluating clustering quality and computing operations process perform profiling with results visualization by the consumption level hardware resources in online mode.

6. References

[1] G. Bo, The Use of Machine Learning Combined with Data Mining Technology in Financial Risk Prevention, Computational Economics, 59 (2022). DOI:10.1007/s10614-021-10101-0.

[2] M. Solanki, Ms. Sharma, A Review of Data Mining Techniques and Its Applications, International Journal of Innovative Research in Computer Science & Technology, 122 (2021) 100-104. DOI:10.55524/ijircst.2021.9.6.23.

[3] C. Peng, Digital Inclusive Finance Data Mining and Model-Driven Analysis of the Impact of Urban-Rural Income Gap, Wireless Communications and Mobile Computing 7 (2022) 1-8. DOI:10.1155/2022/5820145.

[4] N. Rudnichenko, V. Vychuzhanin, I. Petrov, D. Shibaev, Decision Support System for the Machine Learning Methods Selection in Big Data Mining, in: Proceedings of The Third International Workshop on Computer Modeling and Intelligent Systems (CMIS-2020), CEUR-WS, 2608, 2020, pp. 872-885.

[5] V. Vychuzhanin, N. Rudnichenko, Z. Sagova, M. Smieszek, V. Cherniavskyi, A. Golovan, Analysis and structuring diagnostic large volume data of technical condition of complex equipment in transport, in: 24th Slovak-Polish International Scientific Conference on Machine Modelling and Simulations - MMS 2019, Liptovský Ján, Slovakia, 2019. DOI:10.1088/1757-899X/776/1/012049

[6] L. Yunmei, S. Zhang, M. Chen, Y. Wu, Z. Chen, The Sustainable Development of Financial Topic Detection and Trend Prediction by Data Mining, Sustainability, 13 (2021) 7585. DOI:10.3390/su13147585.

[7] X. Li, T. Meng, Enterprise Precision Marketing Effectiveness Model Based on Data Mining Technology, Mobile Information Systems, 11 (2022). DOI:1-10. 10.1155/2022/2020038.

[8] D, Kim, A. Irakoze, Identifying Market Segment for the Assessment of a Price Premium for Green Certified Housing: A Cluster Analysis Approach, Sustainability, 15 (2022) 507-515. DOI: 10.3390/su15010507.

[9] E. Ernawati, S. Baharin, F. Kasmin, A review of data mining methods in RFM-based customer segmentation, Journal of Physics: Conference Series, 2021, pp.1869-1885. DOI:10.1088/1742-6596/1869/1/012085.

[10] D. Saumendra, J. Nayak, Customer Segmentation via Data Mining Techniques: State-of-the-Art Review, 2022. DOI: 10.1007/978-981-16-9447-9_38.

[11] M. Monge, C. Quesada-López, A. Martinez, M. Jenkins, Data Mining and Machine Learning Techniques for Bank Customers Segmentation: A Systematic Mapping Study, 2021. DOI:10.1007/978-3-030-55187-2_48.

[12] I. Petrov, V. Vychuzhanin, N. Rudnichenko, T. Otradskya, Data Mining Information System for Complex Technical Systems Failure Risk Evaluation, in: Proceedings of The Fifth International Workshop on Computer Modeling and Intelligent Systems (CMIS-2022), CEUR-WS, 3137, pp.250-261.

[13] N. Rudnichenko, V. Vychuzhanin, N. Shibaeva, S. Antoshchuk, I. Petrov, Intellectual Information System for Supporting Text Data Rephrasing Processes Based on Deep Learning, in: Proceedings of the 2nd International Workshop on Intelligent Information Technologies & Systems of Information Security (IITSIS), CEUR-WS, 2853, 2021, pp. 228-237.

[14] Y. Wu, L. Liu, Z. Xie, K. Chow, W. Wei, Boosting Ensemble Accuracy by Revisiting Ensemble Diversity Metrics, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp.16469–16477.

[15] F. Rashidi, S. Nejatian, H. Parvin, V. Rezaie, Diversity based cluster weighting in cluster ensemble: an information theory approach, Artificial Intelligence Review, 52 (2019), 1341–1368. DOI:10.1007/s10462-019-09701-y.

[16] R. Burton, S. Cuff, M. Morgan, A. Matt, M. Eberl, GeoWaVe: Geometric median clustering with weighted voting for ensemble clustering of cytometry data, Bioinformatics (Oxford, England), (2022). DOI:39. 10.1093/bioinformatics/btac751.
[17] Z. Wang, H. Parvin, S. Qasem, B. Tuan, K. Pho, Cluster ensemble selection using balanced normalized mutual information, Journal of Intelligent & Fuzzy Systems, 12 (2020) 1–23. DOI: DOI:10.3233/JIFS-191531
[18] S. Bi, W. Liu, Clustering Analysis of Online Teaching Cases and Evaluation of Teaching Results, International Journal of Emerging Technologies in Learning (iJET), 18 (2023) 128-142. DOI:10.3991/ijet.v18i03.38055.
[19] H.Li, X. Ye, A. Imakura, T. Sakurai, LSEC: Large-scale spectral ensemble clustering. Intelligent Data Analysis, 27 (2023) 59-77. DOI:10.3233/IDA-216240.
[20] T. Uçar, A. Karahoca, Benchmarking data mining approaches for traveler segmentation, International Journal of Electrical and Computer Engineering (IJECE), 11 (2021) 409-416. DOI: 10.11591/ijece.v11i1. pp. 409-415.
[21] S, Rogic, L. Kascelan, Segmentation Approach for Athleisure and Performance Sport Retailers Based on Data Mining Techniques, International Journal of E-Services and Mobile Applications, 13 (2021) 71-85. DOI:10.4018/IJESMA.2021070104.
[22] A. Gunandi, H. Awang, E. Alhawad, L. Shabaan, Customer Value and Data Mining in Segmentation Analysis, International Journal of Information Technology and Computer Science Applications, 1 (2023) 20-34. DOI:10.58776/ijitcsa.v1i1.16.
[23] C. González, R. Delgado, J. María, S. Santos, Segmentation of Potential Fraud Taxpayers and Characterization in Personal Income Tax Using Data Mining Techniques, Revista Hacienda Pública Española, 239 (2021) 127-157. DOI:10.7866/HPE-RPE.21.4.4.