Stochastic Game Model of Data Clustering

Petro Kravets[a], Yevhen Burov[a], Oksana Oborska[a], Victoria Vysotska[a], Lyudmyla Dzyubyk[a] and Vasyl Lytvyn[a]

[a] *Lviv Polytechnic National University, S. Bandera Street, 12, Lviv, 79013, Ukraine*

Abstract

A stochastic game model of data clustering under interference conditions is proposed. An adaptive recurrent method and algorithm for stochastic game deciding have developed. Computer simulation of game clustering of noisy data has performed. The parameters influence on the stochastic game method convergence for noisy data clustering is researched. For this purpose, each data point is considered as a separate player with the ability to learn and adapt to the uncertainties of the system. After the selection of clusters is completed, all players calculate the corresponding losses by the criteria of minimizing the total distance between the cluster points formed by the free choice of player strategies. The results obtained are analysed. A stochastic approximation based on the mixed-strategy adjustment method minimizes the mean loss functions on single simplexes.

Keywords

Stochastic game, computer linguistic system, cluster analysis, data clustering, machine learning, mixed strategy, game method, big data, pure strategy, player strategy, natural language processing, stochastic game model, stochastic approximation, clustering method.

1. Introduction

Clustering can accomplished by solving data mining and data visualization problems, grouping and pattern recognition, knowledge extraction and information retrieval, and object classification [1-2]. The purpose of cluster analysis is to find groups of similar objects in a given set or sample. Unlike discriminant analysis, where classes are predefined, cluster analysis determines the composition of clusters [3].

Cluster analysis is used in data science, data mining, natural language processing, machine learning, electronic commerce, information technologies, scientific work, computer linguistics, informatics, document science, pattern recognition, signal processing, marketing, philology, psychology, pedagogy, sociology, medicine, biology, chemistry and other human activity fields to data processing and organization in the classes form for their systematization, grouping and analysis [4].

Cluster analysis in computational linguistics is a tool for creating new classifications, defining models of tokens and finding general patterns of development of whole groups of tokens, for example, to determine propaganda in content for computer linguistic system (CLS) [5].

Cluster analysis allows interlanguage generalizations of the studied concept based on comparison of features possessed by objects in one group. Cluster analysis offers a deeper qualitative analysis of each of the groups. It is necessary to emphasize the role of statistical methods in linguistic research. In the light of the scientific and technological revolution and the growing amount of information (including corpora), linguistics, despite the complexity and versatility of the object of study, requires

IntelITSIS’2021: 2nd International Workshop on Intelligent Information Technologies and Systems of Information Security, March 24–26, 2021, Khmelnytskyi, Ukraine

EMAIL: Petro.O.Kravets@lpnu.ua (P. Kravets); yevhen.v.burov@lpnu.ua (Y. Burov); oksana.v.oborska@lpnu.ua (O. Oborska); victoria.a.vysotska@lpnu.ua (V. Vysotska); liudmyla.v.dziubyk@lpnu.ua (L. Dzyubyk); vasyl.v.lytvyn@lpnu.ua (V. Lytvyn)
ORCID: 0000-0001-8569-423X (P. Kravets); 0000-0001-8653-1520 (Y. Burov); 0000-0002-8647-3305 (O. Oborska); 0000-0001-6417-3689 (V. Vysotska); 0000-0001-6942-9436 (L. Dzyubyk); 0000-0002-9676-0180 (V. Lytvyn)

the development of new methodological foundations of a comprehensive quantitative-systemic approach.

The term cluster is widely used in the exact sciences. In particular, in physics and chemistry, as well as sociology, a cluster characterizes a group of atoms or molecules, a cluster of objects, that is, a form that unites homogeneous particles.

In relation to linguistic objects, a cluster can called a sequence of linguistic elements, such as sounds or parts of speech, as well as a group of dialects or languages that have a number of common features. Thus, noting the organizing function of a cluster as its main characteristic, let us trace which linguistic unities are classified as clusters.

1. An indication of an enlarged linguistic cluster is the genre of the text, since it is characterized by a set of linguistic units of a certain level that implement a single communicative form.
2. The cluster can combine various spelling, syntactic, phonetic, morphological and lexical characteristics of texts.
3. Clusters also include a combination of propositions directly related to lexical groups.

The list goes on.

Clustering process is the objects set division into separate different power subsets depending on their similarity. Selected subsets are called clusters [6]. The elements of a single cluster have common properties [7]. The elements of different clusters differ significantly [8]. Packages of computer programs have developed that implement cluster analysis procedures. Although the existing packages have great power and versatility, but quite complex to use. These are programs like Statistica, SPSS, MatLab, R, and others.

The main groups of cluster analysis tasks:

1. Hypotheses submission based on data research;
2. Hypotheses or research testing for the groups (types) identified in any way determination in the available data;
3. Research of useful conceptual schemes;
4. Development of typology or classification.

The general process of clustering is as follows [9]:

5. The objects characteristics selection and analysing [10];
6. Definition of object metrics [11];
7. Division of multiple objects into clusters [12];
8. Interpretation of clustering results [13].

However, it is always necessary taking into account the shortcomings of cluster analysis.

1. This type of analysis can give unstable clusters. It is very important to pay attention to the preparation of initial data so that the clusters are so to speak clean. However, special attention should be paid to the characteristics based on which the clustering is carried out, and the method that is effective in a particular field of study. The results of the classification should verified by other examples.
2. Cluster analysis implements the inductive method of research from partial to general. Ideally, the sample for classification should be very large, heterogeneous. All hypotheses obtained because cluster analysis must tested. The problem of accurate determination of the number of clusters is also not solved.
3. One of the important problems is the interpretation of clustering results.

In the formulation of clustering problems, the number of clusters may specified or may be unknown a priori [12-14].

2. Related works

The Computer Linguistic System (CLS) is designed to process natural language (NLP) [1]. The main purpose of CLS is the use of artificial intelligence methods, applied linguistics and information technology to understand natural information when performing various tasks both in everyday life and in specialized research (Fig. 1) [1].

Today, the field of computational linguistics is developing rapidly, but most projects are commercialized and one-time. Therefore, there is no-single unambiguous approach, general requirements and recommendations for the design, development and synthesis of relevant CLS. There is also no consensus on the classification of CLS. Thus, according to the classification from Grammarly [1], there are three main types of CLS: analysis, transformation and mixed (Fig. 2). This list is expanded by referral systems, human psychological analysis systems (e.g. IBM Watson on Personality Insights), plagiarism detection (copyright and rewrite) systems, authoring style identification systems, speechless access interface, sign language recognition systems, etc. Stephen William Hawking was one of the most famous people to use a language computer to communicate.

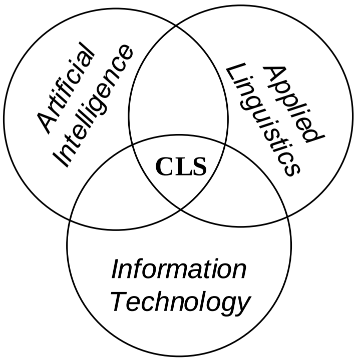

Figure 1: The main directions used for the synthesis of CLS

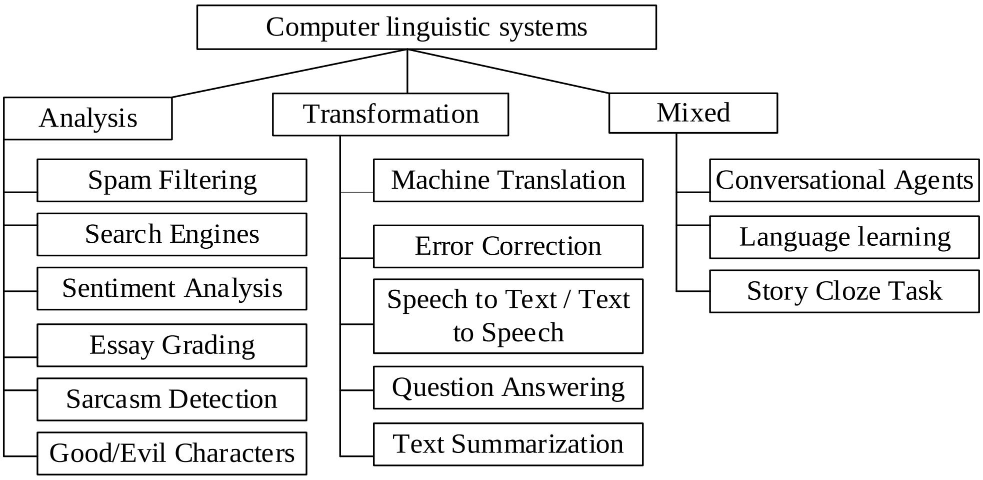

Figure 2: Computer linguistic systems classification

IBM Watson ™ Personality Insights provides an API for retrieving statistics from social networks, corporate data, and other digital communications. The service uses linguistic analytics to infer the internal characteristics of people's personalities through digital communications such as e-mail, text messages, tweets and forum posts.

The main factor in market development and the frequency of implementation of CLS was the motivation to use intelligent devices, cloud solutions and applications based on NLP, which improve customer service in various industries and significantly increase the potential audience of modern information technology users without special knowledge to use them. This is also influenced by the range of tasks to be solved by the different purpose of the CLS (Fig. 1). The main areas of problem solving for CLS are text analysis, text generation, speech recognition and synthesis.

Some of the current tasks belong to some areas, for example, dialog systems rely on such NLP-tools as language recognition, content and context selection, determination of intentions, and then building a dialogue based on the above (ideally - by language synthesis). Thus, a smart assistant must solve the tasks of language recognition, text analysis, text generation and, accordingly, language synthesis. A machine translation solves the problems of text analysis, speech synthesis and text generation. For QA-systems (question-answer) it is enough to solve the problems of text analysis.

Cluster analysis methods are used to solve most problems of computer linguistic systems. Clustering and classification of big text data is now carried out using machine learning. The machine learning software algorithm for processing natural languages in computer linguistics consists of three main parts:

1. Natural language processing [5].
   - tokenization;
   - lemmatization;
   - stop listing;
   - frequency of words;
2. Clustering methods [2-12].
   - TF-IDF;
   - SVD;
   - finding cluster groups;
3. Classification methods - Aylien API [5].

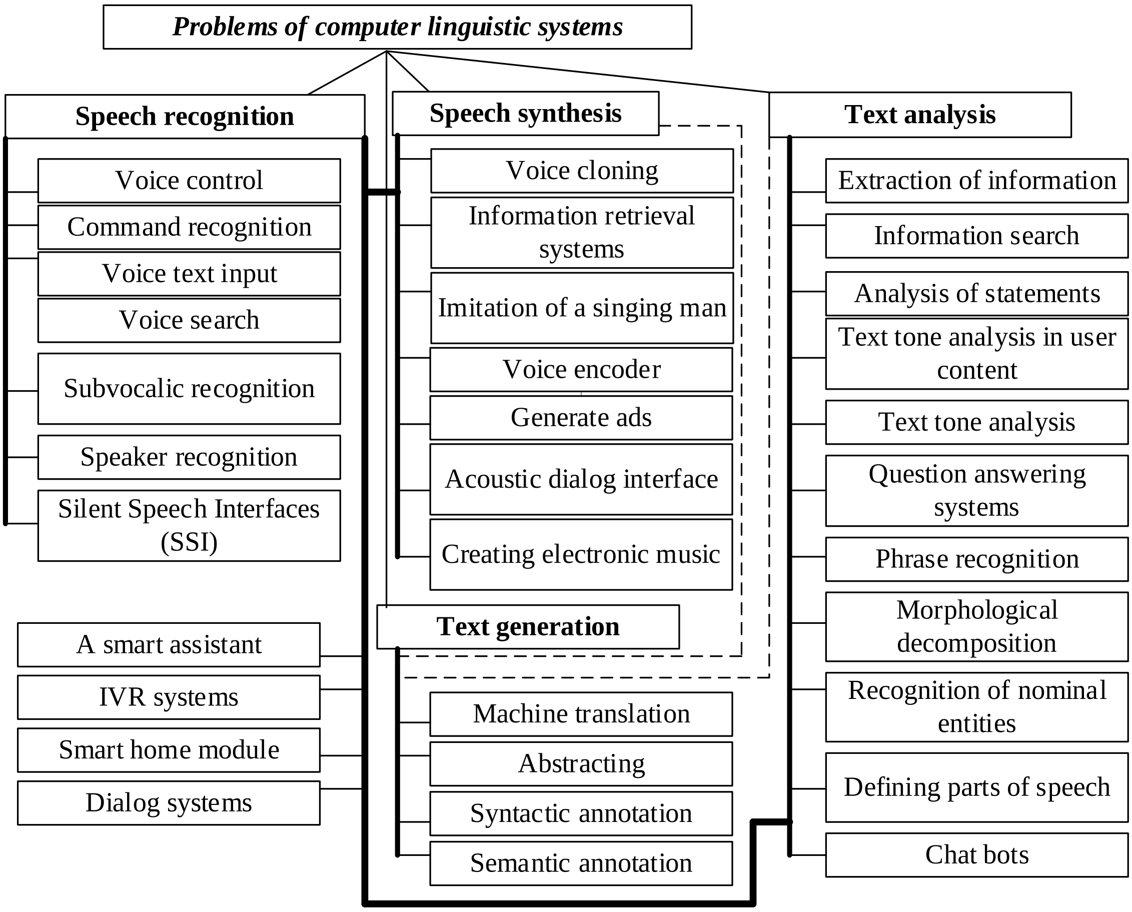

Figure 3: Classification of problems of computer linguistic systems

The task for NLP in computer linguistic systems [1]:

- Develop algorithms to extract features from language;
- Develop algorithms that use the extracted features to solve the broader task.

More Features of linguistic units of measurement analysis in computer linguistic systems [1]:

- Size and arrangement of paragraphs;
- Number of sentences, words, words per sentence, etc.;
- Word position in a sentence;
- Word length;
- Ratio of vowels vs consonants;
- Number of syllables in a word;
- Number of word senses;

- Depth of the word in the dependency tree of the sentence;
- Morphemes: affixes, roots, endings;
- Ngrams;
- Grammatical categories of different POS;
- The word capitalized/hyphenated/compound.

Challenges for computer linguistic systems [5]:

1. Splitting;
2. POS tagging;
3. Parsing;
4. Pragmatics.

Workflow for computer linguistic systems [1]:

1. Research available data and algorithms;
2. Prepare the test set and the baseline;
3. Define metrics;
4. Develop an algorithm:
    - Feature design;
    - NLP pipeline;
    - NLP resources;
    - Approach: rule-based/statistical/machine learning;
5. Implement the solution;
6. Test the solution;
7. Monitor the performance;

Viral headlines identification in computer linguistic systems [1, 5]:

1. Tokenization;
2. Named-entity recognition;
3. Part-of-speech tagging;
4. Ngrams (sequences of elements and their frequencies);
5. Clustering;
6. Machine learning with neuron networks;
7. SentiWordNet.

Approaches to error correction in computer linguistic systems [1, 5]:

1. Pattern matching:
    - Rules make use of:
        a) Coreference resolution;
        b) syntactic parse trees;
        c) POS tags;
        d) Lexical and grammatical dictionaries;
        e) Regular expressions;
        f) Etc.;
    - Rules can be complex, multi-layered;
2. Statistical methods (for example clattering analysis);
3. Machine learning (sometimes pattern matching and simple statistics cannot generalize);
4. Preposition presence and choice:
    - Multiple correct options
        a) Correct but rare usage (I rely mostly {upon=>on} my instinctive feeling);
        b) We have problems {such as/like/with} rapid development;
        c) The economic globalization is of the most concern {by=>to/of/for} each nation;
    - Detection: check every preposition in the sentence;
    - Correction:
        a) Train a classifier (e.g., Random Forest, Logistic Regression);
        b) Prepositions are a closed word class;
    - Needed:
        a) Choose features to use;

        b) Annotated data for training;
    - Ngrams:
        a) Unigrams, bigrams, three-grams…;
        b) Left and right context;
        c) Part-of-speech Ngrams;
    - Grammatical features:
        a) Part of speech;
        b) Dependency relations;
        c) Constituency spans;
    - Semantic features:
        a) Word embeddings;
        b) Semantic role labelling;
        c) VerbNet;
    - Linguistic resources:
        a) Word-form dictionaries;
        b) Governing dictionaries;
    - Meta features:
        a) Dialect (AmE vs BrE);
        b) Genre of the writing;
        c) L1 of the writer;
5. Article presence and choice;
6. Derivational morphology;
7. Run-on sentences;
8. Neural machine translation.
    - Round-trip – translation from English to another language and then back to English;
    - Noisy channel: what if we do translation from bad English to good English?

Machine learning is just used for big data, to reduce labour and material resources. Big data is any data source that has at least one of four common characteristics: volume, diversity, speed, credibility.

In conclusion, we can say that big data and machine learning are closely related to each other, since big data is useless without analysing and extracting information, and machine learning could not coexist without big data, which gives the algorithm experience and learning.

Clustering is a straightforward technique to understand. Objects with similar parameters are grouped together (in a cluster). All objects in a cluster are more similar to each other than to objects in other clusters. Clustering is a type of unsupervised learning because the algorithm itself determines the general characteristics of the elements in the data. The algorithm interprets the parameters that make up each element and then groups them accordingly. Clustering categories [2-14]:

1. K-means method;
2. Density-based spatial clustering for noisy applications - DBSCAN;
3. Clustering algorithm OPTICS;
4. Method of principal components.

However, it is important to note that in clustering, especially in unsupervised learning, the algorithm looks for connections between input data. The beauty of machine learning is finding hidden connections between data, better known as latent connections. For clustering in the search for latent relationships, a model of hidden variables is used, which is applied to study the relationships between the values of variables. The hidden variable model includes [2-14]:

1. EM algorithm;
2. Method of moments;
3. Blind signal separation;
4. Method of principal components;
5. Analysis of independent components;
6. Non-negative matrix decomposition;
7. Singular value decomposition.

Semantic text analysis is one of the key problems of both the artificial intelligence systems creating theory, related to computational linguistics, and natural language processing (NLP). The semantic analysis results can used to solve problems in such areas as, for example [5]:

1. Automatic translation systems;
2. Search engines (The Google search engine is entirely based on semantic analysis);
3. Philology (analysis of copyright texts);
4. Trade (analysis of the demand for certain goods based on comments on this product);
5. Political science (predicting election results);
6. Psychiatry (for diagnosing patients);

Visualization of the results of semantic analysis is an important stage in its implementation, since it can provide fast and effective decision-making based on the analysis results.

Analysis of publications on the network on latent semantic analysis (LSA) shows that the visualization of the analysis results is in the form of a two-coordinate semantic space graph with the words and documents plotted coordinates. Such visualization does not allow unambiguously identifying groups of related documents and assessing the level of their semantic connection by words belonging to the documents. Only cluster labels and centroid coordinates were determined for groups of words and documents without visualization.

Let each object $x \in X$ from many objects $X = (x_1, x_2, ..., x_L)$ be described by the properties vector $x = (x[1], x[2], ..., x[m])$ as quantitative or qualitative object characteristics.

The similarity of the two objects x_i and x_j is determined by the metric of their proximity $D(x_i, x_j)$ in the space of characteristics. The Euclidean distance, Manhattan distance, Chebyshev distance, percentage of discrepancy, Pearson correlation coefficient, etc. are used as metrics [3].

The following methods are most often used to separate many objects into clusters [4, 6]:

1. Hierarchical tree clustering;
2. k-means method;
3. The nearest or farthest neighbor method;
4. Method of unweighted or weighted paired mean;
5. Methods of fuzzy clustering;
6. Use of neural networks;
7. Genetic algorithms;
8. Quenching method.

In general, object clustering can see as the task of optimally dividing objects into groups. The optimization criterion may be to minimize the standard error of cluster allocation:

$$\delta = \sum_{j=1}^{N} \sum_{i=1}^{C_j} \left\| x_i^{(j)} - \overline{x}_j \right\|^2 \rightarrow \min , \qquad (1)$$

where C_j is j-cluster items number;

$\overline{x}_j$ is j- cluster mass centres, a point in the characteristic vectors space with the mean characteristics value for this cluster.

The final stage of cluster analysis is a meaningful interpretation of the clusters formed, during which we identify the factors or cause of grouping of objects into clusters. It should note that different clustering methods could generate different cluster solutions [14]. In addition, the clustering method can detect imported data structures that are not actually present in the analysed data [15]. It is necessary to choose the best clustering methods for the most meaningful decisions in the researched subject area [16]. Experts of relevant subject areas are usually involved to evaluate the quality of clustering.

Based on the results of cluster analysis, you can classify objects, identify conceptual clusters of objects, test and formulate hypotheses about data organization models, compress data by replacing the cluster with its default element, identify the novelty by data that is not included in any of the clusters.

The possibility of software tools of data clustering [17]:

1. Demonstration examples that implement algorithms for real industrial data.
2. Visualization functions that display data in a smaller space (Sammon).

3. Analysis functions designed to evaluate fixed partitions into clusters using index-based methods:
    - Partition index,
    - Xie and Beni's,
    - Alternative Dunn,
    - Dunn, etc.
4. Data clustering methods:
    - GGclust,
    - GKclust,
    - FCMclust,
    - SOM,
    - Hierarchical,
    - PAM,
    - K-medoid,
    - K-means, etc.

The software tools number for data clustering problems solving is developed, in particular:

1. Commercial development:
    - STATISTICA;
    - SPSS Statistics;
2. Popular software packages:
    - Cluster Validity Analysis Platform;
    - Fuzzy Clustering and Data Analysis Toolbox;
    - MatLab.

In practical applications, clustering data typically contains elements of uncertainty [18]. These may be obscure object characteristics, missing object attributes in databases, noisy alerts, and more [19].

In uncertainty conditions are used, in particular [14-19]:

1. Neural networks with training without a teacher;
2. Genetic algorithms;
3. Fuzzy, adaptive clustering methods.

Big data has many characteristics that limit the range of clustering techniques that can work with it. In particular, since big data is a product of modern technologies, methods of their analysis are also at the stage of active development. Thus, the aim of the work is to study statistical methods for clustering large amounts of data with their practical application.

The purpose is a game data-clustering model development with uncertainty elements. For achieve this goal, we need to solve the following tasks:

- Formulate the game clustering data problem,
- An adaptive game method development,
- A computer program model development,
- The obtained results analysing and interpretation.

3. Methods and Materials

Let the set $X = \{x_1, x_2, ..., x_L\}$ be given by the coordinates of the points $x \in R^m$ in m-measurable parametric space. Point coordinates define a normalized characteristic vector for clustering objects. It is necessary N clusters selecting in set $X = \{x_1, x_2, ..., x_L\}$ in particular:

$$\left\{ Y_n, n = 1..N \,\middle|\, \bigcup_{n=1..N} Y_n = X,\ Y_i \bigcap_{i \neq j} Y_j = \varnothing\ \forall (i,j) \in \{1..N\} \right\}$$

with parameters

$$\frac{1}{C_n} \sum_{x \in Y_n} \left\| x_l - x_k \right\| \to \min,\ n = 1..N, \tag{2}$$

where $\|*\| \in R^1$ is Euclidean vector norm;

$C_n = |Y_n|$ is the number of elements included in the cluster Y_n.

In order to find the distribution of the set X on clusters Y_n ($n = 1..N$) is executed by the stochastic game method through a tuple:

$$(I, A^i, \Xi^i \,|\, i \in I), \tag{3}$$

where $A^i = \{a^i(1), ..., a^i(N)\}$ are pure i - player strategies set, which determine the choice of one from the clusters;

$L = |I|$ is I players number;

N is pure strategies i -player;

$\Xi^i : A \to R^1$ is loss function i -player;

$A = \underset{i \in I}{\times} A^i$ is combined player strategies.

The game essence is randomly players moving from one cluster to another. Each player randomly selects a pure strategy in time moments $t = 1, 2, ...$ based on a random event generator $a^i \in A^i$, which determines its entry into the corresponding cluster. Players receive random losses $\xi^i(a)$ with a priori unknown stochastic characteristics after the combined variant $a \in A$ implementation according to (2):

$$\xi_t^i = \frac{1}{C_t^i} \sum_{j \in I} \chi\left(a_t^i = a_t^j\right) \left\| x^i - x^j \right\| + \mu \quad \forall i \in I, \tag{4}$$

where d is dispersion of distribution;

$\mu \sim Normal(0, d)$ is a random variable for the system uncertainty modelling;

$\chi(*) \in \{0,1\}$ is event indicator function;

$C_t^i = \sum_{j \in I} \chi\left(a_t^i = a_t^j\right)$ is the current number of cluster elements with i -player.

The game performance by the medium losses functions is determined:

$$\Xi_t^i = \frac{1}{t} \sum_{\tau=1}^{t} \xi_\tau^i \qquad \forall i \in I. \tag{5}$$

The game purpose is to the medium losses functions system minimization (5) in time:

$$\overline{\lim_{t \to \infty}} \Xi_t^i \to \min \qquad \forall i \in I. \tag{6}$$

Therefore, based on observing the current losses $\{\xi_n^i\}$ every player $i \in I$ must learn how to choose pure strategies $\{a_t^i\}$ so that over time $t = 1, 2, ...$ ensure that the criteria system is fulfilled (6).

The stochastic game solution (2) is performed using adaptive recurrent transformations for mixed strategies vectors. We construct the stochastic game-solving method based on a stochastic approximation of the condition of the complementary rigidity of the deterministic game, which holds for mixed strategies at the Nash equilibrium point [20].

To do this, we define the polyline function of the mean losses of a deterministic game:

$$V^i(p) = \sum_{a \in A} v^i(a) \prod_{j \in I; a^j \in a} p^j(a^j), \; v(a) = M\{\xi_t^i(a)\}. \tag{7}$$

Then the vector condition of complementary slackness will look like:

$$\overset{\rightarrow}{\mathrm{CS}} = \nabla_{p^i} V^i(p) - e^{N_i} V^i(p) = 0 \qquad \forall i \in D, \; e^N = (1_j \,|\, j = 1..N), \tag{8}$$

where $\nabla_{p^i} V^i(p)$ is the gradient of the mean losses function;

$p \in S^M$ is combined mixed player strategies set on a convex single simplex S^M ($M = N^L$).

It is necessary to the additional non-rigidity condition weighing by mixed strategies vectors elements for taking into account the solutions in the single simplex vertices:

$$diag(p^i)(\overset{\rightarrow}{CS}) = 0 \;\; \forall i \in D, \tag{9}$$

where $diag(p^i)$ is square diagonal order matrix N, built of vector elements p^i:

$$diag(p^i)[\nabla_{p^i} V^i - e^{N_i} V^i] = E\{\xi_t^i [e(a_t^i) - p_t^i] \,|\, p_t^i = p^i\}.$$

Recurrent dependence is obtained from (9) based on the stochastic approximation method:

$$p_{t+1}^{i} = \pi_{\varepsilon_{t+1}}^{N} \left\{ p_t^i - \gamma_t \xi_t^i (e(a_t^i) - p_t^i) \right\} \quad \forall i \in I , \qquad (10)$$

where $e(a_t^i)$ is a single vector that indicates a pure strategy choice $a_t^i = a^i \in A^i$;

$\gamma_t > 0$, $\varepsilon_t > 0$ are monotonically declining sequences of positive quantities;

$\pi_{\varepsilon_{t+1}}^{N}$ is projector on single N - measurable simplex S^N [21].

Options γ_t and ε_t determine the convergence conditions of a stochastic game and can be set as follows:

$$\gamma_t = \gamma t^{-\alpha}, \ \varepsilon_t = \varepsilon t^{-\beta}, \qquad (11)$$

where $\gamma > 0$; $\alpha > 0$; $\varepsilon > 0$; $\beta > 0$.

The strategies convergence (10) to optimal values with probability 1 and rms is determined by the γ_t and ε_t parameters ratio [22], which must satisfy the basic stochastic approximation conditions [23-27].

Expandable design ε_t - simplex $S_{\varepsilon_{t+1}}^{N}$ provides the condition

$$p_t^i[j] \geq \varepsilon_t , \ j = 1..N , \qquad (12)$$

that required for statistical information completeness about selected pure strategies. The parameter $\varepsilon_t \to 0$, $t = 1, 2, ...$ is used as an additional element for the recurrence method convergence control.

Pure strategies a_t^i Choosing is performed by players based on dynamic random distributions (10):

$$a_t^i = \left\{ A^i(k) \middle| k = \arg\left(\min_k \sum_{j=1}^{k} p_t^i(a_t^i(j)) > \omega \right), \ k = 1..N \right\} \quad \forall i \in I , \qquad (13)$$

where $\omega \in [0, 1]$ is a real random number with uniform distribution.

Stochastic play begins with untrained mixed vector vectors with element values ($j = 1..N$):

$$p_0^i(j) = 1/N . \qquad (14)$$

The mixed strategies vectors dynamics are determined by the Markov recurrence method (10)-(13).

Each player chooses a pure strategy a_n^i, based on a mixed strategy p_t^i at the moment. This player gets the current loss ξ_t^i for the time $t+1$. Next, this player calculates a mixed strategy p_{t+1}^i according to (10). Method (10)-(13) provides an adaptive pure strategies choice over time by the dynamic mixed strategies reorganization based on the current losses processing [28-35].

The game data clustering quality of is evaluated by:

- The average loss function ($L = |I|$ is the power of many players):

$$\Xi_t = \frac{1}{L} \sum_{i=1}^{L} \Xi_t^i , \qquad (15)$$

- The average norm for mixed player strategies:

$$\Delta_t = \frac{1}{tL} \sum_{\tau=1}^{t} \sum_{i=1}^{L} \left\| p_\tau^i \right\| \qquad (16)$$

Stochastic game solving algorithm is presented:

1. Set the initial parameter values:

$L = |I|$ is number of players;

$t = 0$ is the starting point of time;

$X = \{x_1, x_2, ..., x_L\}$ is a clustering parameters set;

m is number of parameter measurements $x \in R^m$;

N is number of pure player strategies (number of clusters Y_n, $n = 1..N$);

$A^i = \{a^i(1), a^i(2), ..., a^i(N)\}$, $a^i(j) = j$, $i = 1..L$, $j = 1..N$ are pure player strategies vectors;

$p_0^i = (1/N, ..., 1/N)$, $i = 1..L$ is mixed initial strategies of player;

$\gamma > 0$ is the learning step;

ε is simplex parameter;

$\alpha \in (0,1]$ is the learning step order;
$d > 0$ is variance of the plants;
$\beta > 0$ is the speed of expansion of ε - simplex;
$\omega \in [0,1]$ is a valid random value with a uniform distribution.
$t_{\max}$ is the maximum method steps number.

2. Choose action options $a_t^i \in A^i$, $i = 1..L$ according to (13).

3. Get current losses values ξ_t^i, $i = 1..L$ according to (4). The current values of the Gaussian white noise are calculated as:

$$\mu_t = \sqrt{d}\left(\sum_{j=1}^{12} \omega_{j,t} - 6\right). \tag{17}$$

4. Calculate parameter values γ_t, ε_t according to (11).

5. Compute the mixed strategies p_t^i vectors elements according to (10) by $i = 1..L$.

6. Calculate the quality characteristics of clustering Ξ_t (14), Δ_t (15).

7. Set the next point in time $t := t+1$.

8. If $t < t_{\max}$, then go to step 2, otherwise - end.

4. Experiments, Result and Discussions

The stochastic game solvation for data clustering is used by the game method (10)-(13) with parameters:

- $N = 2$,
- $m = 2$,
- $A^i = (1,2)$,
- $t_{\max} = 10^5$,
- $\alpha = 0.01$,
- $\beta = 2$,
- $\gamma = 1$,
- $\varepsilon = 0.999 / N$.

Let two nonempty subsets $Y_1 \cup Y_2 = X$ be visualized within the base set $X = \{Y_1, Y_2\}$. Consider the following three options of a points set organizing for clustering.

Option 1. The subsets do not intersect: $Y_1 \cap Y_2 = \varnothing$. The distance between subsets exceeds the subsets diameters: $S(Y_1,Y_2) > D(Y_2)$ and $S(Y_1,Y_2) > D(Y_1)$, where $D(Z) = \max_{z_1,z_2 \in Z} \|z_1 - z_2\|$ and $S(Y_1,Y_2) = \min_{y_1 \in Y_1, y_2 \in Y_2} \|y_1 - y_2\|$.

The set $X = \{\{(1,3),(3,1),(3,3)\},\{(7,7),(7,9),(9,7)\}\}$ satisfies these conditions. The methods (10)-(13) use provides the stochastic game solution in pure strategies. For this variant of data, the game's solution is subsets that do not intersect: $Y_1 = \{(1,3),(3,1),(3,3)\}$ and $Y_2 = \{(7,7),(7,9),(9,7)\}$.

Fig. 4 shows the graphs of the players' average loss functions on a logarithmic scale Ξ_t and the average rate of mixed strategies Δ_t, that characterize the convergence of a stochastic data clustering game (parameters values α and β must satisfy the basic conditions for stochastic approximation [22]).

The dependence of the average number of steps $\overline{t}$ learning the game from parameter α shown in Fig. 5. Value $\overline{t}$ averaged over $k_{\exp} = 100$ implementations of random processes. The game stopping moment is determined by correct attribution for the set elements X to one of the clusters Y_1 or Y_2 (these clusters are rendered in the set X) and the $\Delta_t \geq 0.99$ approximation condition for the average rate of mixed strategies to 1. The results are obtained for the variance value of the noise $d = 0$.

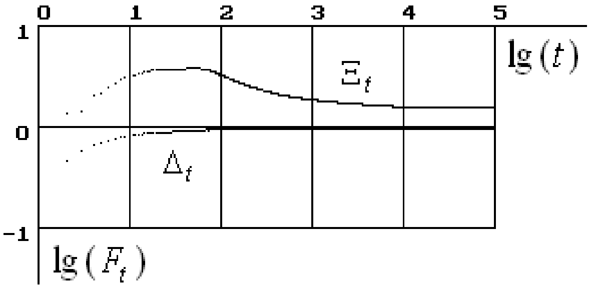

Figure 4: Stochastic game convergence characteristics

For a solvable task, increasing the parameter value α from 0 to 0.7 does not significantly impair the convergence of stochastic play. A significant increase in the average game steps number occurs at $\alpha > 0.7$.

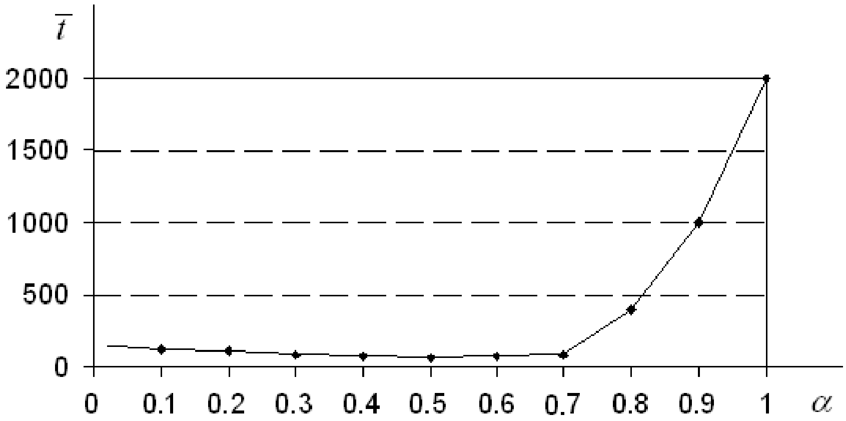

Figure 5: The parameter effect on the game convergence

The stochastic game stability under the noise influence in the white noise form is investigated. The impact of interference dispersion d the average number of steps $\bar{t}$ data clustering game is shown in Fig. 6. The results are obtained for parameter values $\alpha = 0.3$ and $\beta = 2$.

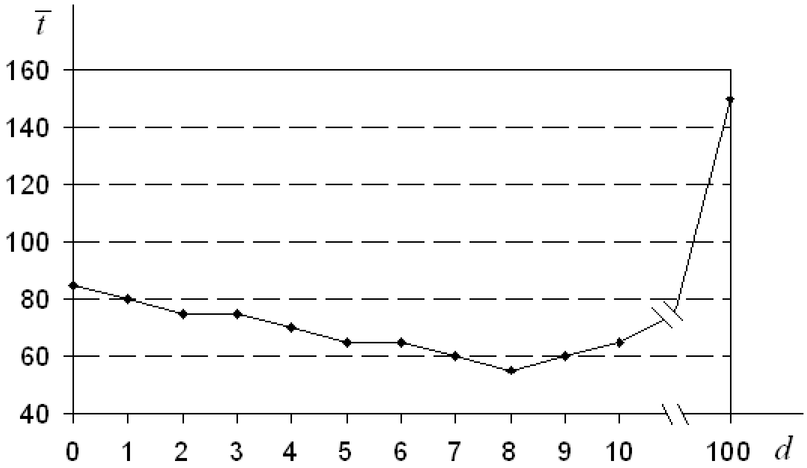

Figure 6: The variance effect on the game convergence

The variance value $d \in [0;50]$ does not significantly affect the data clustering problem solution by the game method (10)-(13). The increase in noise intensity ($d > 50$) leads to a significant increase in the average game steps number required to properly assign elements from the set X to one of the clusters Y_1, Y_2 at the game learning level $\Delta_t \ge 0.99$. The variance limits you set depend on the absolute values of the players' current losses.

As the distance $S(Y_1, Y_2)$ between subsets Y_1 and Y_2 decreases (when the conditions of Option 1 are violated, their boundary elements can be referred to as a subset Y_1 so to the subset Y_2.

Option 2. The subsets intersect: $Y = Y_1 \cap Y_2 \ne \varnothing$.

Points $y \in Y$ in the common subset is placed at the same distance from subsets $Y_2 - Y$ and $Y_1 - Y$:

$$| s(y, Y_1 - Y) - s(y, Y_2 - Y) | < \varepsilon, \quad s(y, Z) = \min_{z \in Z} \|y - z\| \ . \tag{18}$$

The set $X = \{\{ (1,3),(3,1),(5,5)\}, \ \{(5,5),(7,9),(9,7)\}\}$ satisfies those conditions. A point $(5,5) \in Y$ is in the same distance from subsets $Y_2 - Y = \{(7,9),(9,7)\}$ and $Y_1 - Y = \{ (1,3),(3,1)\}$. It can equally be referred to as the cluster Y_1 and to the cluster Y_2. For given inputs, methods (10)-(13) ensure that stochastic games are solved in pure strategies. The possible solutions are:

- $Y_1 = \{(1,3),(3,1)\}$, $Y_2 = \{(5,5),(7,9),(9,7)\}$.
- $Y_1 = \{(1,3),(3,1),(5,5)\}$, $Y_2 = \{(7,9),(9,7)\}$.

Option 3. In the set X no subsets are rendered Y_1 and Y_2: $X = Y_1 = Y_2$.

Let $X = \{(4,6),(5,5),(6,4)\}$ and can divided into clusters according to criteria (6). The possible solutions at $N = 2$ are: $Y_1 = \{(4,6)\}$, $Y_2 = \{(5,5),(6,4)\}$, $Y_1 = \{(4,6),(5,5)\}$, $Y_2 = \{(6,4)\}$.

Method (10)-(13) provides a solution to the clustering data game problem in mixed strategies when $0 < | X | \le 2$. Fig. 7 shows the convergence characteristics of a stochastic division game $X = \{(4,6),(6,4)\}$ to $N = 2$ clusters. The game settings are as follows: $\alpha = 0.3$, $\beta = 2$, $d = 0$.

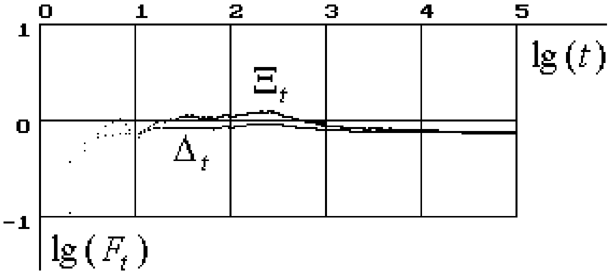

Figure 7: Stochastic game convergence 2×2 characteristics

The average rate function for mixed strategies Δ_t does not reach the logarithmic zero. This indicating that the possible solutions for the mixed strategies are:

- $Y_1 = \{(6,4)\}$, $Y_2 = \{(4,6)\}$.
- $Y_1 = \{(4,6)\}$, $Y_2 = \{(6,4)\}$.
- $Y_1 = \varnothing$, $Y_2 = \{(4,6),(6,4)\}$.
- $Y_1 = \{(4,6),(6,4)\}$, $Y_2 = \varnothing$.

The set X power and the player’s number increasing leads to a decrease in the convergence rate of the stochastic game, which is manifested in the increase in the steps number required to cluster data.

Fig. 8 shows a graph of the average number for stochastic game learning steps from the input data number. The results are obtained for the following game method parameter values: $\alpha = 0.3$, $\beta = 2$, $d = 0$, $N = 2$.

The data intended for clustering are obtained accidentally by the normal law of the points coordinates distribution on a plane. Two clusters of points with normal distribution parameters are generated

$$Normal(E\{(10,10)\}, d(9)), \ Normal(E\{(5,5)\}, d(9)) \ . \tag{19}$$

Moment $\bar{t}$ completion of the game is determined by the condition $\Delta_t \geq 0.99$. The results obtained are averaged over $k_{\exp} = 100$ experiments.

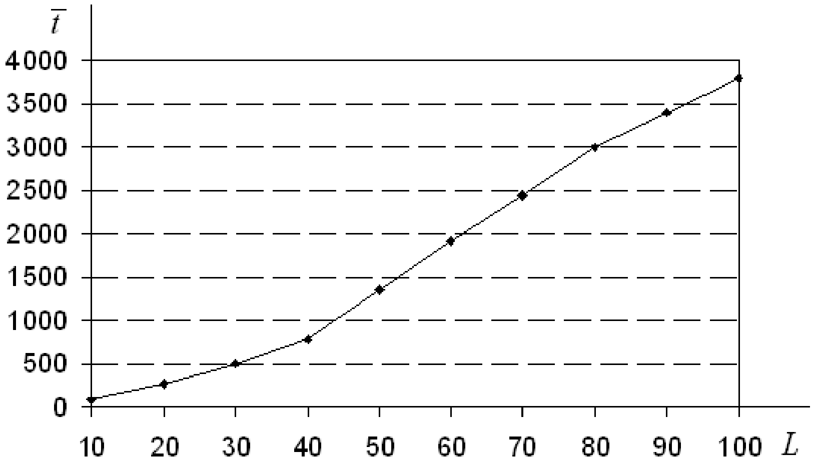

Figure 8: The average game steps number dependence from clustering points number

As the point's number allocated for clustering increases, the steps number required to stochastic game learning to divide data into clusters increases. Achieving acceptable for practical applications characteristics of the stochastic game convergence is determined by fine-tuning the parameters of the game method within the basic relations provided by the stochastic approximation theory [22].

5. Conclusion

This article proposes a new data clustering method based on the stochastic game theory results.

A stochastic game model of data clustering under interference conditions is proposed. An adaptive recurrent method and algorithm for stochastic game deciding have developed. Computer simulation of game clustering of noisy data has performed. The parameters influence on the stochastic game method convergence for noisy data clustering is researched. The results obtained are analysed.

The developed and researched game method (10)-(13) solves the clustering noisy data problem. For this purpose, each data point is considered as a separate player with the ability to learn and adapt to the uncertainties of the system. The net player's strategies are to choose one of a fixed clusters number.

After the selection of clusters is completed, all players calculate the corresponding losses by the criteria of minimizing the total distance between the cluster points formed by the free choice of player strategies.

The resulting losses are used by players to rearrange the dynamic vectors of mixed strategies, which form the basis of a random mechanism for generating pure player strategies. A stochastic approximation based on the mixed-strategy adjustment method (10) minimizes the mean loss functions on single simplexes. Unresolved in this paper is the question of autonomous determination of the clusters number in solving the stochastic game of clustering noisy data.

6. References

[1] M. Romanyshyn, Intro to Natural Language Processing, Grammarly, Inc., 2017.

[2] P. Trebuňa, J. Halčinová, Experimental modelling of the cluster analysis processes, volume 48 of Procedia Engineering, 2012, pp. 673-678. doi: 10.1016/j.proeng.2012.09.569

[3] M. Z. Hossain, M. N. Akhtar, R. B. Ahmad, M. Rahman, A dynamic K-means clustering for data mining, volume 13(2) of Indonesian Journal of Electrical Engineering and Computer Science, 2019, pp. 521-526. doi: 10.11591/ijeecs.v13.i2. pp. 521-526

[4] P. Kravets, Game method for coalitions formation in multi-agent systems, volume 1 of Computer Sciences and Information Technologies, 2018, pp. 1-4. doi: 10.1109/STC-CSIT.2018.8526610

[5] V.-A. Oliinyk, V. Vysotska, Y. Burov, K. Mykich, V. Basto-Fernandes, Propaganda Detection in Text Data Based on NLP and Machine Learning, volume Vol-2631 of CEUR workshop proceedings, 2020, pp. 132-144.

[6] A. S. Shirkhorshidi, S. Aghabozorgi, T. Y. Wah, T. Herawan, Big data clustering: a review, volume 8583 of Lecture Notes in Computer Science, 2014, 707-720. doi: 10.1007/978-3-319-09156-3_49

[7] P. Kravets, V. Lytvyn, V. Vysotska, Y. Ryshkovets, S. Vyshemyrska, S. Smailova, Dynamic Coordination of Strategies for Multi-agent Systems, volume 1246 of Advances in Intelligent Systems and Computing, 2021, pp. 653-670, doi: 10.1007/978-3-030-54215-3_42

[8] P. Kravets, V. Pasichnyk, N. Kunanets, N. Veretennikova, O. Husak, Adaptive Strategies in the Multi-agent "Predator-Prey" Models, volume 1247 of Advances in Intelligent Systems and Computing, 2021, pp. 285-295. doi: 10.1007/978-3-030-55506-1_26

[9] R. Bordawekar, B. Bandyopadhyay, O. Shmueli, Cognitive database: A step towards endowing relational databases with artificial intelligence capabilities, arXiv:1712.07199, 2017.

[10] Y. Oktar, M. Turkan, A review of sparsity-based clustering methods, volume 148 of Signal processing, 2018, pp. 20-30. doi: 10.1016/j.sigpro.2018.02.010

[11] G. Komaki, E. Teymourian, V. Kayvanfar, Z. Booyavi, Improved discrete cuckoo optimization algorithm for the three-stage assembly flowshop scheduling problem, volume 105 of Computers & Industrial Engineering, 2017, pp. 158-173. doi: 10.1016/j.cie.2017.01.006

[12] R.J. Kuo, S.S. Chen, W.C. Cheng, C.Y. Tsai, Integration of artificial immune network and K-means for cluster analysis, volume 40(3) of Knowledge and information systems, 2014, pp. 541-557. doi: 10.1007/s10115-013-0649-3

[13] A. Shirazinia, S. Chatterjee, M. Skoglund, Channel-optimized vector quantizer design for compressed sensing measurements, in Proceedings of International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 4648-4652. doi: 10.1109/ICASSP.2013.6638541

[14] C. Sarbu, Fuzzy Clustering of Environmental Data in NOVA Science Publishers, 2015, pp. 33-56. URL: http://real.mtak.hu/25464/1/KHch1inCurrApplChemometr1_13.pdf#page=49

[15] S. Askari, N. Montazerin, A high-order multi-variable fuzzy time series forecasting algorithm based on fuzzy clustering, volume 42 of Expert Systems with Applications, 2015, pp. 2121-2135. doi: 10.1016/j.eswa.2014.09.036

[16] D. Mustafi, G. Sahoo, A hybrid approach using genetic algorithm and the differential evolution heuristic for enhanced initialization of the k-means algorithm with applications in text clustering, volume 23(15) of Soft Computing, 2019, pp. 6361-6378. doi: 10.1007/s00500-018-3289-4

[17] O. Kilinc, I. Uysal, Learning latent representations in neural networks for clustering through pseudo supervision and graph-based activity regularization, arXiv:1802.03063, 2018.

[18] Z. Hu, Y. Bodyanskiy, O. Tyshchenko, V. Tkachov, Fuzzy clustering data arrays with omitted observations, volume 11(6) of International Journal of Intelligent Systems and Applications, 2017, pp. 24-32. doi: 10.5815/ijisa.2017.06.03.

[19] M. Elidrisi, N. Johnson, M. Gini, J. Crandall, Fast adaptive learning in repeated stochastic games by game abstraction, in Proceedings of the International conference on Autonomous agents and multi-agent systems, 2014, pp. 1141-1148.

[20] E. N. Barron, Game theory: an introduction, in John Wiley & Sons, 2013.

[21] L. Wang, Z. Wang, S. Liu, An effective multivariate time series classification approach using echo state network and adaptive differential evolution algorithm, volume 43 of Expert Systems with Applications, 2016, pp. 237-249. doi: 10.1016/j.eswa.2015.08.055

[22] H. J. Kushner, D. S. Clark, Stochastic approximation methods for constrained and unconstrained systems, volume 26 of Springer Science & Business Media,2012.doi:10.1007/978-1-4684-9352-8

[23] G. Le Ray, P. Pinson, Online adaptive clustering algorithm for load profiling, volume 17 of Sustainable Energy, Grids and Networks, 2019, p. 100181. doi: 10.1016/j.segan.2018.100181

[24] V. Kiyko, V. Lytvyn, L. Chyrun, S. Vyshemyrska, I. Lurie, M. Hrubel, Forest Cover Type Classification Based on Environment Characteristics and Machine Learning Technology, volume

1158 of Communications in Computer and Information Science, 2020, pp. 501-524. doi: 10.1007/978-3-030-61656-4_34

[25] V. Danylyk, V. Vysotska, V. Lytvyn, S. Vyshemyrska, I. Lurie, M. Luchkevych, Detecting Items with the Biggest Weight Based on Neural Network and Machine Learning Methods, volume 1158 of Communications in Computer and Information Science, 2020, pp. 383-396. doi: 10.1007/978-3-030-61656-4_26

[26] A. Gozhyj, I. Kalinina, V. Gozhyj, V. Danilov, Approach for Modeling Search Web-Services Based on Color Petri Nets, volume 1158 of Communications in Computer and Information Science, 2020, pp. 525-538. doi: 10.1007/978-3-030-61656-4_35

[27] A. Gozhyj, I. Kalinina, V. Gozhyj, Fuzzy cognitive analysis and modeling of water quality, in: International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, 2017, pp. 289-294. doi: 10.1109/IDAACS.2017.8095092

[28] P. Bidyuk, A. Gozhyj, I. Kalinina, V. Vysotska, Methods for Forecasting Nonlinear Non-stationary Processes in Machine Learning, volume 1158 of Communications in Computer and Information Science, 2020, pp. 470-485. doi: 10.1007/978-3-030-61656-4_32

[29] I. Lurie, V. Lytvynenko, S. Olszewski, M. Voronenko, A. Kornelyuk, U. Zhunissova, O. Boskin, The Use of Inductive Methods to Identify Subtypes of Glioblastomas in Gene Clustering, volume Vol-2631 of CEUR Workshop Proceedings, 2020, pp. 406-418.

[30] V. Lytvynenko, I. Lurie, J. Krejci, M. Voronenko, N. Savina, M. A. Taif, Two Step Density-Based Object-Inductive Clustering Algorithm, volume Vol-2386 of CEUR Workshop Proceedings, 2019, pp. 117-135.

[31] S. Mashtalir, O. Mikhnova, M. Stolbovyi, Multidimensional sequence clustering with adaptive iterative dynamic time warping, volume 18 of International Journal of Computing, 2019, 53-59. doi: 10.47839/ijc.18.1.1273

[32] R.J. Kosarevych, B.P. Rusyn, V.V. Korniy, T.I. Kerod, Image Segmentation Based on the Evaluation of the Tendency of Image Elements to form Clusters with the Help of Point Field Characteristics, volume 51(5) of Cybernetics and Systems Analysis, 2015, pp. 704-713. doi: 10.1007/s10559-015-9762-5

[33] S. Babichev, An Evaluation of the Information Technology of Gene Expression Profiles Processing Stability for Different Levels of Noise Components, volume 3 of Data, 2018, art. 48. doi: 10.3390/data3040048

[34] S. Babichev, B. Durnyak, I. Pikh, V. Senkivskyy, An Evaluation of the Objective Clustering Inductive Technology Effectiveness Implemented Using Density-Based and Agglomerative Hierarchical Clustering Algorithms, volume 1020 of Advances in Intelligent Systems and Computing, 2020, pp. 532-553. doi: 10.1007/978-3-030-26474-1_37

[35] S. Babichev, M.A. Taif, V. Lytvynenko, V. Osypenko, Criterial analysis of gene expression sequences to create the objective clustering inductive technology, in: Proceedings of the International Conference on Electronics and Nanotechnology, ELNANO, 2017, pp. 244-248. doi: 10.1109/ELNANO.2017.7939756