DOI: https://doi.org/10.15276/aait.01.2020.1
UDC 004.9

Methodology of information monitoring and diagnostics of objects represented by quantitative estimates based on cluster analysis

Nataliia O. Komleva[1)]
ORCID: https://orcid.org/http://orcid.org/0000-0001-9627-8530, komleva@opu.ua, Scopus ID: 57191858904
Vira V. Liubchenko[2)]
ORCID: https://orcid.org/ http://orcid.org/0000-0002-4611-7832 lvv@opu.ua, Scopus ID: 56667638800
Svitlana L. Zinovatna[1)]
ORCID: http://orcid.org/0000-0002-9190-6486, zinovatnaya.svetlana@opu.ua, Scopus ID: 57206667710
[1)] Odesa National Polytechnic University, 1, Shevchenko Avenue. Odesa, 65044, Ukraine
[2)]Hochschule für Angewandte Wissenschaften Hamburg, Fakultät Life Sciences, Ulmenliet 20 Hamburg, 21033, Germany

ABSTRACT

The paper discusses the methodological foundations of informational diagnostics on the base of cluster analysis for the objects represented by quantitative estimates. The literature review showed that the application of cluster analysis in some cases was successful; also, the theory of cluster analysis is well developed, and the properties of methods and distance measures are studied, which indicates the appropriateness of using the cluster analysis apparatus. Therefore, the development of a general methodology to diagnose any objects represented by quantitative estimates is a topical task. The purpose of this work is to develop methodological bases for determining diagnostic states and behavioral patterns for objects represented by quantitative estimates on the base of cluster analysis. Because of informational diagnostics is a targeted activity on the assessment of object state based on a dynamic information model, the model of a diagnosis object is discussed first. We examine the lifecycle of instances of diagnosis objects that are described by a plurality of parameters whose values are determined by a time slice along the lifeline of the instance. It is shown that a different number of measured values characterize each state of the diagnosis object. There are identified characteristics that should be analyzed to indicate a threat to the instance and the need for supportive procedures to prevent premature interruption of an instance's lifecycle. Experts should carry out the formalization of conditions for termination of the life cycle of the diagnosis object and formation of the list of supporting procedures. Because the quality of any information technology depends on the input data quality, a procedure for the analysis of diagnostic characters is developed. In order to start the diagnosis as early as possible and apply the available data as fully as possible, the methodologies for one-, two- and N-step diagnosis are developed. All procedures used cluster order. Transition patterns are defined for the two-step diagnosis, as well as trend patterns are defined for the N-step diagnosis. Transition patterns allow diagnosing the improvement, worsening, or stability of the diagnosis object state. The procedure for the diagnostic characters analysis and the methodologies of diagnosis is new scientific results. The application of the developed methodologies is demonstrated in the example of diagnosing students' success. In this case, the curriculum provides the domain model. Examples of diagnosing states and behavior, as well as identifying recommended reactions, are provided. For one-step diagnostics, the presence of the influence of the latent factor and the diagnostic signs that show significant instability are investigated. For one- and two-step diagnostics, the conditions for forming a risk segment are provided.

Keywords: Informational Diagnostics; Cluster Analysis; Diagnostic Character; Pattern; Trend

***For citation*: Nataliia O. Komleva, Vira V. Liubchenko, Svitlana L. Zinovatna. Methodology of Information Monitoring and Diagnostics of Objects Represented by Quantitative Estimates Based on Cluster Analysis. *Applied Aspects of Information Technology.* 2020; Vol.3 No.1: 376–392. DOI: https://doi.org/10.15276/aait.01.2020.1**

INTRODUCTION

The availability of large amounts of data allows a person to increase the amount and quality of information that is used as a basis for making various decisions. A separate area of research – information analytics – is dedicated to the application of essential analytics functions to provide decision-makers with the information they need to make the right decisions. Usually, they try to solve two types of problems:

1) identify the causes that led to the facts being recorded;

2) predict the behavior of the objects that the data is collected for.

In this paper, we will consider another problem – the problem of informational diagnostics. Informational diagnostics is a targeted activity on the assessment of diagnosis object state based on a dynamic information model with predetermined similarity criteria [1].

Examples of situations in which informational diagnostics are needed are ongoing staff assessment of the required competencies using the 360-degree method, student achievement monitoring in the semester, etc.

An object of any nature can be diagnosed if it is variable over time, and change data is available. Among the important features of the information diagnostic procedure are the following:

1) not the object itself, but the available data about it is analyzed;

2) the integrity of the object data collection, which is determined by the object's domain model, must be ensured for successful diagnosis;

3) the technological chain of diagnostic procedures should allow diagnostics of an unfamiliar object with reproducible object evaluation results.

There are a number of tasks in which the diagnosis object (DO) is relevant only to its current status (the applicant's points obtained during the introductory campaign for certain disciplines determine the list of higher education institutions available for the budgetary form of education; the specific job position determines the possibility of the first interview, the data on the technical inspection of the vehicle determine the set of repair work, etc.).

However, at the same time, the process of development of an object in time, its behavior, and the nature of manifestation of peculiarities in different situations may be of interest. Any object - material or abstract – has a life cycle of its development, consisting of a sequence of processes, from the conception of the formation of the object to the completion of its existence. Often, not only the entire development life cycle of an object falls into the area of attention, but its fragment, for example, a software project from the moment of formulation of requirements to the first stable release, an employee from the first to the last day of internship, a student during training at an educational institution. If it is possible to identify specific times and highlight data about an object that is bound to those moments, then you can consider the task of diagnosing the behavior of the object over time.

Thus, depending on the tasks being solved – static or dynamic – the methodology of diagnosis must operate with the concepts of "state" or "behavior" of the object. If the number of time points for which object data can be retrieved is a finite sequence, the diagnostic process is a multi-step process.

LITERATURE REVIEW

The term diagnostics came from medicine and concerned about the identification and classification of health problems through research and evaluation. Over time, medical institutions have accumulated large volumes of patient data, and some of the diagnostic work has become possible only through observation data. Accordingly, researchers turned to solve the problem of interpreting the results of clinical testing using algorithms for statistical processing and data analysis. For example, the study [2] proposes a system to interpret clinical examination results for the diagnosis of coronary heart disease based on the decision tree algorithm. The paper [3] provides a survey of current techniques of knowledge discovery in databases using data mining techniques that are in use in today's medical research, particularly in heart disease prediction. In [4], the authors described how to improve evaluation of the reliability of the diagnostic in medicine based on conformal predictors that allow carrying out a probabilistic classification.

Data-driven diagnosis techniques have been transferred to other domains. The paper [5] provides a state-of-the-art review of approaches to using multivariate statistical tools for characterization normal variations and detection the abnormal changes in the industrial process. Also, statistical diagnosis methods support fault detection and fault identification for rigorous analysis. Paper [6] presents a data fusion approach for machinery fault diagnosis using fuzzy measures and fuzzy integrals, which consists of a feature-level data fusion model and a decision-level data fusion model. The fuzzy c-means analysis method was employed to identify the relations between a feature set and a fault prototype. This paper [7] looks at the use of data-driven models built for monitoring, fault diagnosis, optimization, and control. Particular attention is paid to latent variable models because they provide reduced dimensional models for high dimensional processes.

The review paper [8] gives a full picture of fault detection and diagnosis in complex systems from the perspective of data processing. Such a system is a data-processing system based on information redundancy, in which the data and human's understanding of the data are two fundamental elements. Human's understanding may be an explicit input-output model representing the relationship among the system's variables. Therefore, the traditional data-driven fault diagnosis methods rely on the features extracted by experts. The feature extraction process is exhausting work and dramatically affects the result. The paper [9] demonstrates that deep learning provides an effective way to extract the features of raw data automatically and eliminate the effect of handcrafted features.

One of the essential tools of diagnosis is cluster analysis because detection of the current state of an object can be considered as its classification into a group of objects that are similar in certain features.

For example, an article [10] describes the clustering of students according to their behavior in the online learning process; it is shown that this approach has the potential to provide more adaptive tools within the intellectual learning system or the teacher-person.

Obviously, in general, determining the status of a diagnosed entity is the information that is extracted from the data in order to simplify or improve the decision-making process. For example, diagnosing student behavior in the learning process allows timely intervention in the learning process. "Intervention has long been practiced in higher education to provide assistance for at-risk or underachieving learners. With the development of learning analytics, the delivery of intervention has

been informed by data-driven approaches to identify learners' problems and provide them with just-in-time and personalized support. However, the intervention has been claimed to be the greatest challenge in learning analytics and has yet to be widely implemented" [11]. In this case, "when identifying at-risk students, it is important to minimize false negative (i.e., type II) error while not increasing false positive (i.e., type I) error significantly" [12]. Also, according to the authors [13], "for learning analytics research to account for the diverse ways technology is adopted and applied in course-specific contexts".

Along with the status classification, it is essential to classify the behavior of the object or system. In [14], the authors argue that "classification of the behavior of users ... provides an understanding of people's sequence of activities within a period of time" and "is of great interest to scientific communities". In particular, it allows studying specific patterns of behavior and identify the causes that led to them. In [15], it is shown by the example of technical objects that "comparison of the models created in different time points estimates the deterioration in a condition of the object caused by the long operation".

Article [16] provides an overview of different clustering methods, including the problem of data clustering and the definition of terms. In particular, the algorithms of hierarchical clustering and partitional clustering are discussed; the k-means algorithm, the fuzzy c-means algorithm, the Expectation-Maximization algorithm, and others are described in detail. [17] emphasized that "the validity and usefulness of the output of different clustering methods can only be evaluated by the user in the context of each particular application".

Although there are many methods of cluster analysis that are widely used and whose properties have been well studied, the theory of cluster analysis continues to evolve. In particular, existing clustering methods are being refined; for example, in [18], an advanced hierarchical clustering algorithm is proposed, in which "the feature weights are cluster dependent, allowing a feature to have different degrees of relevance at different clusters". Also of interest is the extension of the capabilities of existing algorithms, for example, [19] describes the extension of the DyClee algorithm for categorical data, "approaches, global and local density, to generate clusters have been modified to capture categorical data".

In addition, the methods based on modern information technologies are being developed. For example, the survey [20] focused on using deep neural networks to learn a clustering-friendly representation, resulting in a significant increase in clustering performance. In chapters of [21], authors discuss not only deep learning for clustering but also blockchain data clustering, cybersecurity applications, scalable distributed clustering methods for massive volumes of data, clustering big data streams.

It should take into account that the value of clustering results depends directly on the quality of the data on which it is executed. Therefore, it is crucial to pre-process the data to clear the original data and control the data [22]. In [23], the problem of anomaly detection in the initial data was considered, and the necessity of a systematic approach to the characterization of anomalies was substantiated. The example of the time series shows how this facilitates the use of the k-means algorithm in combination with hierarchical clustering.

Considering that real processes are characterized by massive data sets, much attention is paid to the selection of relevant and informative variables for clustering [24,25]. In the absence of expert information, special algorithms are used to select informative features, including Wrappers, Filters, and Embedded. Wrapping and filtering algorithms create subsets of informative attributes, using a search in the space of possible input data, and then wrapping algorithms evaluate the received subset of inputs by learning from the available data for the full model and filtering for the less complicated model. Built-in algorithms use learning heuristics to evaluate the importance of input features. In the presence of expert advice on the selection of informative features, particular attention should be paid to dealing with large unstructured datasets; lack of experience in using the Occam shaver for the sake of minimizing computation time can lead to a loss of meaning [26].

K-type clustering algorithms are widely used in real-world applications, such as marketing research and data processing, for clustering very large datasets, due to their efficiency and ability to handle numerical and categorical variables that are ubiquitous in real databases [27-28]. The main problem with using k-means algorithms in data processing is the choice of variables since these algorithms cannot automatically select variables, but handle all variables equally in the clustering process. In practice, the selection of variables for the clustering task is often made based on an understanding of the business problem and the data used. Therefore, in the presence of experts who can determine the set of such variables, it is very advisable to use the k-means algorithm.

It should be noted that the choice of the cluster analysis method is influenced by the nature of determining the degree of similarity between different objects [29]. Prior information about how to control the weights of these objects, depending on the estimation vectors and the external conditions of diagnosis, allows selecting the most appropriate metrics for determining the distance between objects

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

[30]. The Euclidean distance is most commonly used, but if more massive objects are more distant from each other, the square of the Euclidean distance is used. If it is necessary to reduce the impact of large individual deviations (single emissions), it is advisable to use the Manhattan distance [31, 32]. If it is necessary to take into account the characteristic with the largest different only, the distance Chebyshev is used [33]. The influence of the degree of object distinction on such a characteristic can be controlled by using the Minkowski distance [34]; this is done using the parameters of the stepwise weighting of distances by individual coordinates and progressive weighting of distances between objects by particular characteristics.

Therefore, we can conclude that it is advisable to use cluster analysis as a tool for informational diagnosis. Because the solution to the problem of cluster analysis is based on datasets characterizing clustered objects, its application will not require the extraction of additional data for the DO. As a result, DOs are divided into groups of similar objects, enabling them to generalize descriptions of their states and behavior, as well as provide them with a diagnostic interpretation based on the analysis of aggregated observations of the respective cluster.

THE PURPOSE OF THE ARTICLE

Let there be a DO that has many instances $E=\{e_1, \ldots, e_i, \ldots, e_M\}$. The state of an object can be described by a vector of estimates $MS=(m_1,\ldots,m_i,\ldots,m_K)$, $\forall i=1,K$. Over time, the object changes its status and the essential characteristics of the object may change over time. Therefore, the vectors of estimates at different points in time may be different in the semantics and values of the estimates, as well as in the number of coordinates of the vector. Measurement of the characters of the object is performed at discrete times. The condition and behavior of objects should be monitored to reduce the probability of problems in their functioning with some corrective actions.

The aim of this work is to develop methodological bases for determining diagnostic states and behavioral patterns for diagnosis object, described by quantitative traits, using cluster analysis.

THE DIAGNOSIS OBJECT

Each instance of DO is characterized by a lifetime in the system from moment t_0 to moment t_f. Instances of an object may appear on the system at different times:

$$t_0(e_i) \neq t_0(e_j). \quad (1)$$

Accordingly, life expectancy Δt for different objects may be different:

$$\Delta t(e_i) = t_f(e_i) - t_0(e_i). \quad (2)$$

The lifetime may be scheduled, but it may stop when certain conditions are created. The lifetime is discrete, i.e., each time slice is associated with the state of the object instance $S_l(e_i)$.

Limit values are defined for each evaluation. In the general case, the power of the set *MS* may be different for different states and different instances.

The time interval $<t_{l-1},t_l>$ cannot always be broken down into smaller intervals, since the intermediate time slice may not be able to obtain the value of m_p, or the value obtained will correspond only to part of m_p.

Therefore, mp can be calculated as a result of the aggregation of intermediate values or obtained at time t_l without depending on the intermediate values:

$$m_p = \begin{cases} A_{e=1}^{nt}(mp_{te}), \\ m_t \end{cases}, \quad (3)$$

where: *A* is an aggregation function; *nt* is the number of intermediate values; mp_{te} is the value of the measure at an intermediate point in time t_e from the interval $<t_{l-1},t_l>$; m_t is the absolute value of the measure at time t_l without aggregation of intermediate values.

The purpose of diagnosis is to prevent premature termination of the life cycle of an instance of an object.

An instance of a diagnosis object (IDO) may terminate ahead of schedule if the aggregate value for all measures of the $S_l(e_i)$ state is less than the threshold value δ_{lim}^1 or the number of values of the $S_l(e_i)$, state measures that exceed the critical m_c value exceeds the limit value δ_{lim}^2:

$$A_{e=1}^{|MS|}(m_e) < \delta_{lim}^1 \text{ Ú } |\{m_e: m_e < m_c\}| > \delta_{lim}^2. \quad (4)$$

Each measure, in addition to a certain quantitative value of m_p, is also defined by a set of different dimensions:

$$v_p = <m_p, D>, \quad (5)$$

where: *D* represents a set of measurement values, each of which has a specific value for v_p.

Factors of the analysis complexity are that:

– different IDOs may have different numbers of prior states at the time of analysis;

– each state can be characterized by a different number of measure values.

THE MAIN STAGES OF THE DEVELOPMENT OF DIAGNOSTIC TOOLS

Consider, in general terms, the process of developing and verifying means of information diagnosis, depending on the number of "steps" – moments of time at which data can be obtained about the object (Fig. 1). We will analyze the cases for one, two, and *N* steps.

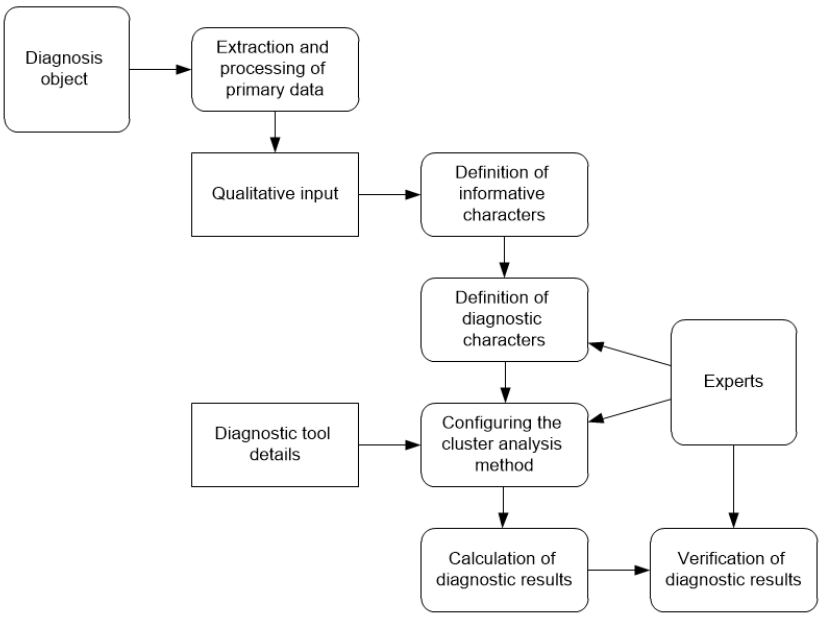

Fig. 1. **The process of informational diagnosis based on cluster analysis**
***Source*: compiled by the author**

Each of the diagnostic tools is constructed to study the status and/or behavior of the DO. Obviously, DO is of interest in a specific context that is determined by the subject area. Therefore, the domain and the role of DO in it determine the set of diagnosis characters. Thus, the diagnosis is performed not for the DO, but for the DO models, and the quality of the diagnosis results depends on the adequacy of the DO model used. The adequacy of the DO model will mean the correct qualitative and quantitative description of the DO by the selected set of diagnosis characters with a given degree of accuracy. The quantitative assessment of adequacy is determined according to the objectives of the study and the significance of the characteristics of the model [35].

All the IDOs studied are described by primary data, which should reflect their basic properties. If this condition is not met, the resulting DO model cannot be considered adequate. In the process of primary data collection, synchronization of data over time, compliance with the required data types (generally text, numbers, video, and audio) and elimination of irrelevant data must be ensured in this process [36].

Once the primary data is received, it must be processed to clear, normalize, eliminate gaps, and duplicates.

At this stage, the data types are checked for compliance with the requirements of the subsequent processing algorithms. In case of discrepancy, an attempt is made to bring the data to the desired type: moving from real numbers to integers or from real numbers to higher accuracy to lower numbers by rounding, "packing" numeric and text data into XML or JSON formats, moving from table representation to non-relational, such as translating SQL constructs into MongoDB counterparts and the like. If the required data conversion is not possible, the values are cleared.

If data collection conditions varied from object to object and the values of their properties were measured on different scales, data normalization should be performed. Note that the scope of this work can be reduced by strengthening data collection requirements.

Gap processing is an important operation because in some cases, data gaps can indicate that data collection conditions are not met, which has led to data loss. The decision on the possibility of further use of the IDO described in the data with omissions. It is advisable to make an expert way: you can either start procedures to recover lost data or exclude IDO from the sample. In other cases, omissions in the data may indicate that the IDO is actually outside the scope of the domain and was previously mistaken for it. In this case, the IDO is clearly excluded from consideration.

The presence of duplicates in the data can be associated with systematic errors (constant or described by law, processes that occur in the interaction of the tool and the object of measurement, etc.), random errors (accidental obstacles in the DO, the means of measuring their

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

properties, etc.) and gross errors (failures of measuring equipment, errors of personnel, unexpected changes of conditions of carrying out measurements of DO properties, etc.).

As a result of the processing of the primary data, qualitative input is obtained. On their basis, informative features are determined - those data that allow obtaining the values of diagnostic features by means of mathematical calculations and theoretical and logical conclusions. Effective identification of informative features allows reducing the dimension of the diagnosis task. Non-informative information is used to set up the diagnosis method. In the future, such data should be weeded out as irrelevant.

When defining a diagnosis task, a set of diagnostic features is determined by an expert or a group of experts in the field. For example, in determining the success of students diagnostic signs are grades from exams and coursework, in the analysis of the process of software development - the number of errors in the program code and overdue terms of delivery of stages of work, in the diagnosis of the car – the values of key indicators that are responsible for structural parameters technical condition obtained directly by direct measurement or indirect measurement when installing a diagnostic device (stand). If certain diagnostic features have the ability to identify those that are not self-contained but are aggregated in higher-order features, then experts should recommend removing these features from the diagnostic kit to reduce the dimension of the diagnosis task.

The method of cluster analysis as a diagnostic tool allows for solving a number of tasks.

1. Determination of static patterns – DO states – and dynamic patterns – transitions between DO states. In this case, the solution requires the choice of a clustering algorithm that allows you to reconcile the computational complexity of the diagnostic process, as well as the form of the cluster (arbitrary shape, hypersphere), input (number of clusters, distance threshold for truncation hierarchy, degree of fuzzy, etc.) and results (cluster centers) membership matrix, tree structure or binary cluster tree, etc.). In the presence of expert assumptions about static and dynamic patterns, it is possible to compare them with the data obtained.

2. Determination of the properties of a certain IDO in the framework of complete reproduction of the conditions of its functioning environment. This makes it possible to determine the identity of an IDO to one of the clusters that do not intersect.

3. Detection of systematic displacements of values of DO properties under the influence of latent factors. The presence of such shifts means the need to expand the set of diagnostic features by a more detailed study of the nature of DO and conditions of its functioning.

4. Getting the probability of assigning the IDO to each of the clusters according to the selected feature of the distance calculation between the objects. This allows calculate and expertly analyze the entropy of the source of the properties of the sample IDO.

5. Involvement of certain means of measuring the properties of DO and checking the appropriateness of their use. This allows you to get an optimal set of measurement tools and formalize the conditions for their use.

Before performing the cluster analysis, you need to set the settings related to the selected tool, namely:

– for two- and *N*-step diagnostics to determine the lifetime of IDO in the system and the set of measures according to (1) – (4);

– for a one-step diagnosis, which is a separate case of *N*-step, determine the properties for the condition $S_l(e_i)$.

You must also select the cluster analysis method that will be used. As shown earlier, each of the cluster analysis methods has its own characteristics, which must be considered depending on the diagnostic tool used (one-, two- or *N*-step).

The obtained diagnostic data are subject to verification. At this stage, expert evaluation of the results is made, and conclusions are drawn as to their correctness and expediency of using the selected cluster analysis methods, measuring instruments, and the conditions of the DO operation. Subject matter experts formulate recommendations and observations aimed at improving information diagnosis procedures.

The already established methodology is used for diagnosing new objects (Fig. 2).

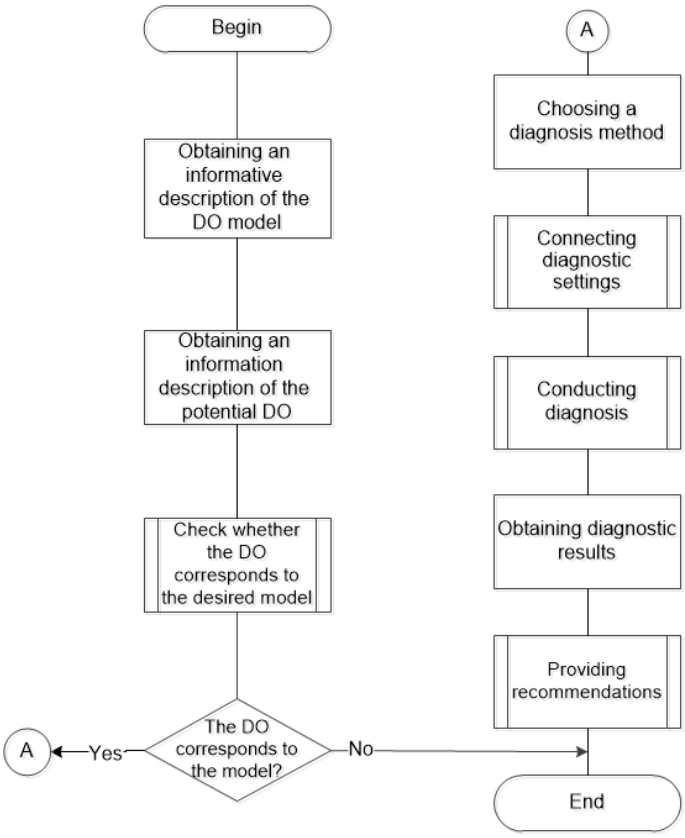

Fig. 2. **Flow-chart of the algorithm of the complete cycle of the study of DO**

***Source*: compiled by the author**

Information on the nature, principles, and conditions of operation of potential DOs allows the identification of DO models.

The negative result of such verification may be the case:

– the absence of a well-established diagnostic technique for the received DO model: recommended action – revision (if possible) of diagnosis characters;

– DO / DO model mismatch: recommended action – re-check DO model with gross errors (e.g., personnel errors);

– poor (incomplete) DO description: recommended action – engage experts to adjust descriptions.

If the DO model is adequate, then the desired diagnostic technique is selected, and all diagnostic procedures are performed. If there are recommendations in the system settings, they are provided based on the DO results.

THE ANALYSIS OF DIAGNOSTIC FEATURES

Before you begin the diagnosis procedure, it is necessary to confirm the quality of characters according to which it will be performed. After all, the peculiarity of the analyzed situation consists in the fact that the evaluations describing the behavior of the object reflect not only the characters of the object but also the characteristics of the evaluator and the influence of external factors. Therefore, it is desirable to determine whether one or more characters are under the influence of the latent factor.

On the basis of experience, it can be stated that the influence of the latent factor will be reflected in the estimates by the systematic shift of the values of the estimates on a particular character. For example, launching a new, well-publicized product will lower the product sales from the same niche; grades given by a too demanding teacher will be lower than the other grades, etc. Therefore, a character that is influenced by the latent factor can be identified by identifying estimates that are different from many other estimates.

The computational complexity of the analysis of all sample observations for the shift in estimates will increase with the increasing number of observations, which is not a computationally effective solution.

Therefore, a solution with constant complexity would be appropriate. Classification is performed to diagnose object states. It results not only in splitting cluster observations but also in the calculation of the coordinates of the centers for each of the clusters. Therefore, centroid coordinates for each feature

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

represent the average value estimates on this basis of observations included in the respective cluster. Accordingly, it can be argued that the analysis of the coordinates of the cluster centers may reveal a shift of estimates due to the influence of the latent factor.

The analysis procedure consists of six steps.

1. Bringing all estimates of one measurement interval for the absolute value assessments not to affect the influence of the character while clustering.

2. Executing clustering observations using a known method of clustering.

3. Calculating the spare values of the b mean m and mean-square deviation σ for the coordinates of each center of clusters.

4. Determining confidence intervals for the coordinates of the centers of clusters. As the diagnostics of the characters are of a recommendation nature and is performed on a small number of samples, it is possible to limit ourselves to rough estimates of the intervals $m\pm\sigma$ or $m\pm2\sigma$.

5. Checking whether going beyond the confidence intervals for one coordinate occurs in all centers of clusters

6. In case of a steady shift, analyzing possible reasons for its occurrence and deciding as to accounting estimates on the respective character for further diagnostics.

The computational complexity of steps 1−2 depends on the sample size. However, their result - the clustering of observations − is used in diagnostic techniques. Therefore, in contrast to the solution with the analysis of all observations in order to detect a systematic shift, the computational complexity of the actual analysis of diagnostic characters is negligible.

ONE-STEP DIAGNOSING

One-step diagnosing corresponds to the situation in which observation results are used to determine the state of the object. In the case of applying clustering as a diagnostic tool, the state of the object is determined according to the cluster, which is described by the coordinates of the center. Such an interpretation will be adequate only if all IDO's are more or less equidistant from the center of the cluster. The major disadvantage of most distance measures is that they use averaging. Because of this, IDOs, which are very different from the center in one or a small number of characters and are close to the center in other characters, can get into the cluster.

Therefore, there is a need to find such objects, the state of which cannot be diagnosed on the basis of cluster center analysis. It is necessary to find objects whose estimations go beyond the confidence interval $m\pm2\sigma$ for at least one coordinate to identify abnormal cases.

The diagnostic procedure for a sample IDO consists of four steps.

1. Executing clustering for IDO's.

2. For each cluster.

a. Calculating the value of standard deviation σ for each coordinate center.

b. Detecting IDO's beyond the confidence interval $m\pm2\sigma$ for at least one coordinate, removing them from the cluster, and adding them to a set of abnormal cases.

3. Diagnosing the state of clustered IDO's based on the information about the centers of clusters.

4. Diagnosing states of IDO's from a set of abnormal cases.

As a result of diagnostic, the set of clusters are obtained that definitely contain IDOs from item 3 and, by expert decision, may include IDOs from item 4.

If you need to diagnose new IDO's which satisfy the DO model and have sufficient data for diagnostics, you need to solve the following problems:

a) whether there is a limit to the number of items in the sample;

b) whether the new IDO has significant differences in its operating conditions, composition, etc.

If there are no restrictions on the number of elements, or the addition of a new element IDOs not leads to an excess of the number, that is, the computational complexity of the problem is acceptable, the new IDO is added to the samples and re-clustered according to the method described above. In another case, a new IDO is diagnosed according to the method of classification by the minimum distance between the IDO and the centers of clusters in accordance with the applied measure of the distance between objects.

If the new IDO differs from those already diagnosed objects with properties that are not included in a model, it is appropriate to include it in the sample and re-execute clustering. This will allow you to track possible changes in the clusters and decide whether to include these properties in the DO model. If it is not possible to add new items because of the number limit, the new IDO is added to the sample with the exclusion of an already existing item.

TWO-STEP DIAGNOSING

In the case when the IDO estimates and the clustering results for them obtained for two consecutive time points' $t-1$ and t are available, the possibility to diagnose changes appears (Fig. 3).

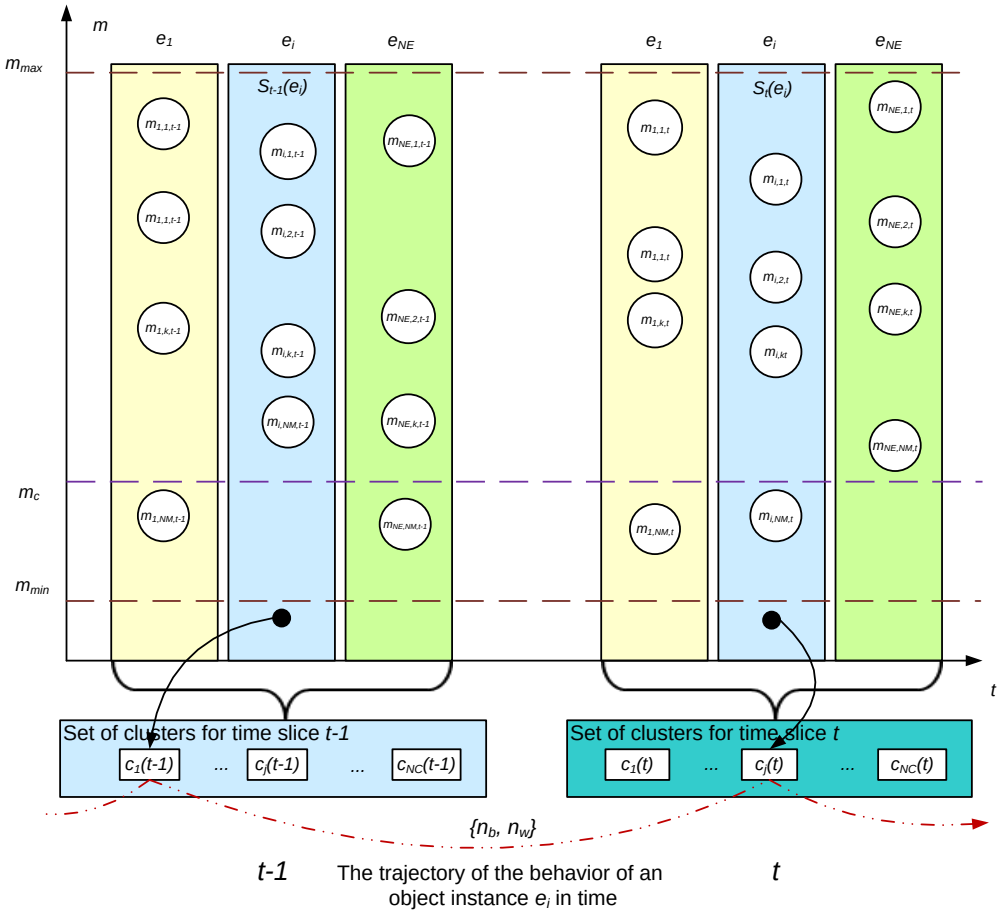

Fig. 3. **Scheme of formation of the behavioral pattern of IDO**

***Source*: compiled by the author**

For each cluster, the appropriate level of quality of IDO's which it contains can be determined, so that the order relation can be determined on clusters. Then, we can define three transition patterns:

a) at moments $t-1$ and t the object falls into clusters of the same order;

b) at time t the object worsens its position compared to the time $t-1$;

c) at time t the object improves its position compared to the time $t-1$.

The degree of risk associated with a worsening condition and the opportunities associated with an improvement in condition will vary depending on the severity of the deterioration or improvement. It is obvious, for example, that the transition $1\rightarrow2$ is better than the transition $1\rightarrow3$. To account for this difference, we introduce appropriate quantitative indicators.

If the order relation is entered on the clusters, the best cluster is number 1, and with the degradation, the number of clusters increases by one, then

– the degradation intensity is calculated as

$$n_w = \begin{cases} \#Cl_t - \#Cl_{t-1} \text{ if } \#Cl_t > \#Cl_{t-1}, \\ 0 \text{ otherwise} \end{cases}$$

where $\#Cl_{t-1}$ and $\#Cl_t$ are the numbers of the clusters to which the DO at times $t-1$ and t respectively;

– the improvement intensity is calculated as

$$n_b = \begin{cases} \#Cl_{t-1} - \#Cl_t \text{ if} \#Cl_{t-1} > \#Cl_t \\ 0 \text{ otherwise} \end{cases}.$$

Now the difference we gave as an example will be reflected in the fact that for the first case $n_w=1$ and for the second case $n_w=2$.

The diagnostic procedure consists of five steps.

1. Executing clustering for IDO at moments $t-1$ and t. The sets of characters at moments $t-1$ and t may differ.

2. Establishing an order relation on clusters detected at moments $t-1$ and t.

3. Determining a behavioral pattern and calculating the corresponding intensity for each EDI.

4. Analyzing the reason of deterioration for IDO's that experienced it. These objects form a risk group and require observation at the next iteration.

5. Analyzing the cause for the IDO's that improve their position. These objects form a group of possibilities and need support in the next iteration.

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

It is obvious that IDO's that are not in risk or possibility groups should not be neglected. However, they demonstrate stability, and a one-step diagnosis is sufficient to determine their strategy.

N-STEP DIAGNOSING

In cases when estimates of objects made at more than two points in time are available, diagnosing can be expanded by studying behavioral trends.

Suppose that we have available data on N observations, therefore, it is possible to perform N clusterings of observations and determine the N–1 transition pattern for each IDO.

Three diagnostic indices can be considered:

– state constancy index

$$R_m = \frac{N_m}{N-1},$$

where: N_m is the number of transitions without changing the state;

– state improvement index

$$R_b = \frac{N_b}{N-1};$$

where: N_b is total improvement intensity on all IDO transitions with the improvement of the states;

– state degradation index

$$R_w = \frac{N_w}{N-1},$$

where: N_w is the total degradation intensity at all IDO transitions with the degradation of states.

We can now describe the procedure of the *N*-step diagnosing.

1. Executing clustering for IDO's in each of the N points in time. The sets of characters may differ at different times.

2. Establishing an order relation on clusters identified for each of the N points in time.

3. Determining the number of transitions without changing state N_m, the total degradation intensity N_b, and total state intensity N_w for each object.

4. Computing values of indices R_m, R_b, R_w for each object.

5. If R_m value is greater than a predetermined threshold value, then the behavior of the object can be considered constant. In this case, one-step diagnosing based on the last-moment data should be applied.

6. If R_b value is greater than a predetermined threshold, then the object is prone to improving its state. For such objects, it is necessary to create conditions that will not significantly interfere with their work.

7. If R_w value is greater than a predetermined threshold, then the object is prone to degradation. For such objects, it is necessary to analyze the causes of such behavior and provide them with observations in work.

CASE STUDY

To evaluate the proposed procedures, the success status of students was diagnosed. We used the data of 47 students studying on a bachelor's program on Software Engineering. As diagnostic characters, we used the grades received on exams and term papers. Analysis of the distribution of exam grades and test grades confirmed the hypothesis that they were statistically similar. Therefore, in order to reduce the dimension space of characteristics, it was decided not to include test grades in the model.

Thus, the diagnosis object in the case is the student; the instances of the diagnosis object are concrete students. Because the purpose of the study is a specific set of competencies and learning outcomes, the curriculum may be used as the domain model [37]. Accordingly, the success of each student is characterized by the grades he has received for the courses. It is natural to monitor the learning performance at the examinations to get regularly the descriptions of objects instances with comparable characteristics.

Let us start by analyzing the diagnostic characters. Because the evaluation is performed on a 100-point scale, there is no need to reduce the scores to a single interval. A well-known k-means algorithm, whose parameter is the number of clusters, is used to perform clustering. In our case, the number of clusters was set as k=4, which corresponds to the traditional classification of students in accordance with their learning success. The results of calculating the cluster centers for the first four semesters are shown in Table 1. The designation $m_{i.j}$ represents the grade obtained in a jth course in the ith semester, m_i and σ_i represent the mean and standard deviation of cluster centers corresponded to the ith semester.

Highlighted values of cluster centers in Table 1 go beyond the intervals $m \pm \sigma$. There is no sign of a systematic shift for any diagnosis characters. However, attention should be paid to $m_{2.6}$ and $m_{4.1}$, which tend toward overestimation, as well as to m1.3, which demonstrates significant instability.

Table 1. **Data for the analysis of diagnostic characters**

Variab-les	Cluster centers			
	#1	#2	#3	#4
m1.1	94,2	92,0	**88,9**	**66,0**
m1.2	93,6	95,7	**78,9**	69,0
m1.3	**96,6**	**70,0**	82,1	**77,0**
m1.4	92,5	90,0	83,6	72,4
*m*1	94,2	86,9	83,4	71,1
σ1	1,7	11,5	4,2	4,7
m2.1	**80,7**	69,1	86,5	**62,5**
m2.2	93,5	76,4	72,7	**89,2**
m2.3	86,1	70,3	**87,0**	76,5
m2.4	96,2	74,6	**61,9**	71,8
m2.5	93,6	84,5	80,0	**63,3**
m2.6	**98,1**	**91,6**	74,6	**88,0**
*m*2	91,4	77,8	77,1	75,2
σ2	6,6	8,7	9,5	11,6
m3.1	**93,9**	**88,8**	63,4	65,8
m3.2	90,7	77,4	65,6	65,0
m3.3	**88,4**	77,8	81,7	**63,3**
m3.4	92,6	76,4	82,8	65,3
*m*3	91,4	80,1	73,4	64,9
σ3	2,4	5,8	10,3	1,1
m4.1	**97,0**	**87,8**	**84,6**	65,6
m4.2	95,6	72,5	82,0	75,0
m4.3	**92,5**	68,8	75,4	**64,0**
m4.4	93,3	72,0	69,4	**78,1**
m4.5	95,4	**89,0**	**67,4**	73,0
*m*4	94,8	78,0	75,8	71,1
σ4	1,8	9,6	7,5	6,1

***Source*: compiled by the author**

The use of one-step diagnostics we consider in the example of the second semester (Table 2).

Table 2. **Distribution of students between clusters**

Cluster	Number of students
#1	17
#2	11
#3	13
#4	6

***Source*: compiled by the author**

To diagnose students' states, information about cluster centers (see Table 1) should be analyzed. For example, cluster #1 brought together a group of A-students, but they received good grades from the two courses. The connection between both courses and the future specialty usually is not evident for students. Therefore, it can be diagnosed that a group of A-students is trying to reduce the time spent by reducing attention to "unimportant" courses. Accordingly, corrective action by a higher education institution should focus on informing this group of students about the value of learning outcomes of different courses.

Consider the definition of abnormal cases in the example of cluster #4. We already know (see Table 1) that in the second semester, each object is represented by six grades. Data for cluster #4 are collected in Table. 3.

Table 3. **Data on students belonged to cluster #4**

Student	$m_{2.1}$	$m_{2.2}$	$m_{2.3}$	$m_{2.4}$	$m_{2.5}$	$m_{2.6}$
7	60	100	93	90	60	90
12	60	80	75	61	65	85
13	60	75	73	67	60	85
20	60	85	82	92	60	98
25	60	95	76	61	75	95
28	**75**	100	60	60	60	75
m	62,5	89,2	76,5	71,8	63,3	88,0
σ	6,1	10,7	10,9	15,1	6,1	8,2

***Source*: compiled by the author**

Therefore, cluster #4 includes one abnormal case. Because student #28 in one grades differ from the corresponding coordinate of the center by more than 2σ, he should be assigned to abnormal cases.

Let us turn to two-step diagnostics. The order of the clusters is set according to the mean value mi calculated for the coordinates of the centers, which are shown in Table 1. Actually, data about the centers of the clusters in Table 1 are already sorted by the order. To illustrate the application of the two-step diagnostics, we consider the results for the transition between first and second examinations (Table 4).

Table 4. **Analyzed transition patterns**

Pattern	Number of students	Diagnose	Power
$1 \to 1$	13	Stable	
$1 \to 2$	1	Risk	n_w=1
$1 \to 3$	3	Risk	n_w=2
$1 \to 4$	2	Risk	n_w=3
$2 \to 1$	1	Opportunity	n_b=1
$2 \to 2$	1	Stable	
$2 \to 3$	1	Risk	n_w=1
$2 \to 4$	0	Risk	n_w=2
$3 \to 1$	3	Opportunity	n_b=2
$3 \to 2$	6	Opportunity	n_b=1
$3 \to 3$	8	Stable	
$3 \to 4$	3	Risk	n_w=1
$4 \to 1$	0	Opportunity	n_b=3
$4 \to 2$	3	Opportunity	n_b=2
$4 \to 3$	1	Opportunity	n_b=1
$4 \to 4$	1	Stable	

***Source*: compiled by the author**

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

As we can see, taking into account previous data provides the possibility for analyzing behavior patterns. As a result of the analysis, we can divide the students set into three subsets: those who degraded their states, those who have maintained their states, and those who have improved their states. Thus, for students who have changed their states, the number of diagnostic information increases.

As an example, let us return to cluster #1, which gathered A-students in the second semester. Among the 17 cluster members (see Table 2), 76 % maintained their state, 6 % improved their state by one position, and 18 % improved their state by two positions. We can conclude that most students retain their performance. However, some students seek to improve their academic performance. Therefore, when planning teaching activities, it is necessary to pay attention to motivational activities.

Now we turn to multi-step diagnostics, which is considered by the example of clustering for observations of four consecutive examinations. A piece of behavioral analysis results demonstrated by students is shown in Table 5. The obtained values of R_m, R_b, and R_w are informational characteristics of the trends that characterize the behavior of students and need further study.

Table 5. **Results of 4-step diagnosis**

Trend	R_m	R_b	R_w
$1 \to 1 \to 1 \to 1$	1	0	0
$1 \to 1 \to 2 \to 1$	0,3	0,3	0,3
$1 \to 2 \to 2 \to 2$	0,7	0	0,3
$1 \to 3 \to 1 \to 1$	0,3	0,7	0,7
$1 \to 3 \to 2 \to 3$	0	0,3	1
$1 \to 3 \to 2 \to 4$	0	0,3	1,3
$1 \to 4 \to 3 \to 1$	0	1	1
$1 \to 4 \to 4 \to 4$	0,7	0	1
…			

***Source*: compiled by the author**

Studying the properties of identified trends with two-step diagnostics allows determining the required level of intervention in the learning process. Let limit the intervention need by the cases when the student degrades his state with time, that is, the transition between the time moments t and t-1 is characterized by n_w>0. The number of transitions between clusters for the studied sample is shown in Table. 6.

Table 6. **Number of transitions between clusters**

Type of transition	Semesters		
	1-2	2-3	3-4
n_w=n_b=0	23	22	25
n_w = 1	5	9	6
n_w = 2	3	3	3
n_w = 3	2	0	0
n_b > 0	14	13	13

***Source*: compiled by the author**

The transitions without changing the state give the highest contribution (about 50 %) in each semester (Fig. 4). Significant degradation (with n_w=3) is diagnosed between the first and second semesters only. The contribution of these cases is about 4 %; the behavior of students who get into this segment of risk should be carefully analyzed with the mandatory activation of appropriate corrective pedagogical practices. The behavior of students, who demonstrate degradation with n_w=1 and n_w=2 also needs to be analyzed, but this problem has a lower priority, in particular under resource restriction.

Particular attention should be paid to stable transitions with n_w=n_b=0. The absence of a change in cluster number means that two-step diagnostics can be reduced to one-step diagnostics, and it is possible to form a risk segment based on cluster numbers. The pie charts show the percentage of contributions for the 1→1, 2→2, 3→3, and 4→4 inter-semester transitions between clusters. In our case, the tendency to remain in the worst cluster #4 over time slightly increases – from 4 % to 12 %. An expert should analyze the behavior of all students who get into this segment of risk and give the recommendations of appropriate corrective pedagogical practices.

Additionally, the gathering of data about trends for a variety of cases provides the possibility in the future to involve the machine learning tools for diagnostics and prediction of behavioral patterns. It will give the possibility to reduce the requirement to the quality of objects description.

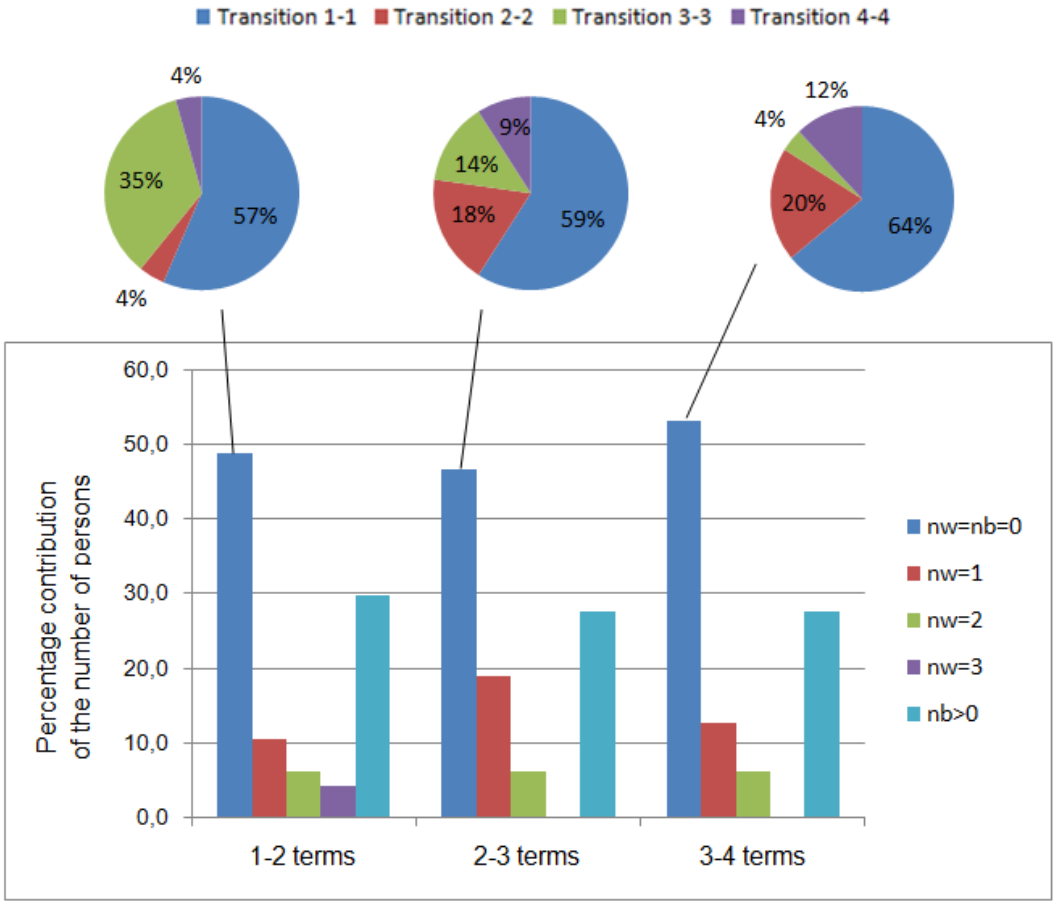

Fig. 4. **Percentage of contributions of different types of transitions of state**
***Source*: compiled by the author**

CONCLUSIONS

An analysis of the features of the application of cluster analysis methods as a diagnostic tool showed that diagnostic characters of diagnosis objects should be used quantitative estimations. At the same time, it is possible to configure the diagnostic technique by determining the similarity measures in the cluster analysis method.

In the paper, we built the model of the diagnosis object. It includes the object lifetime in the system, the set of measures with limit values and corresponding measurements, and the conditions for early termination of the diagnosis object existence.

The main stages in the development of tools for informational diagnostic based on cluster analysis are identified. The lists of tasks that can be solved with the cluster analysis method as a diagnostic tool and the appropriate settings for diagnosing are given.

Analysis of the measures that describe the behavior of an object allows detecting the presence of latent factors leading to a systematic shift in the values of measures for a particular character or group of characters. The procedure of measures analysis is formalized as well as its computational complexity is determined.

One-step diagnosis is the basic one; the procedure for a one-step diagnosis enables one to form the clusters of similar objects as well as to detect the abnormal cases. The features of new objects diagnosing, depending on their similarity to the objects in the sample, are given.

Two-step diagnosis allows determining the patterns of objects transition from one cluster to another as time goes by. The two-step diagnosis procedure is developed, and the degradation and improvement measures for the diagnosis object behavior are introduced.

The developed procedure of the *N*-step diagnosis determines the transitions trends. The indexes of stability, improvement, and degradation are proposed as trend informational characteristics.

Experimental confirmation of the performance of diagnostic procedures was obtained for data about students' success. Examples of diagnosing the states and behavior of students, as well as possible

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

reactions to the results of the diagnosis are demonstrated.

The proposed procedures allow diagnosing the risk of premature life interruption for instances of diagnosis objects represented by quantitative estimates. It is possible due to calculating the powers of degradation and improving the state, as well as forming the patterns and trends of transitions. This result is essential for different tasks because it improves the quality of work with diagnosis objects in the application domains.

REFERENCES

1. Marasanov, V., Sharko, A. & Stepanchikov, D. "Model of the Operator Dynamic Process of Acoustic Emission Occurrence While of Mate-rials deforming". In: Lytvynenko, V., Babi-chev, S., Wójcik, W., Vynokurova, O., Vyshe-myrskaya, S., Radetskaya, S. (eds). *Lecture Notes in Computa-tional Intelligence and Decision Making. ISDMCI 2019. Advances in Intelligent Systems and Computing*. 2020; Vol. 020: 48–64. DOI: https://doi.org/10.1007/978-3-030-26474-1_4.

2. Wiharto, W., Kusnanto, H. & Herianto, H. "Interpretation of Clinical Data Based on C4.5 Algorithm for the Diagnosis of Coronary Heart Disease". *Healthcare Informatics Research.* 2016; Vol. 22(3): 186–195. DOI: https://doi.org/10.4258/hir.2016.22.3.186.

3. Soni, J., Ansari, U., Sharma, D. & Soni, S. "Predictive Data Mining for Medical". Diagnosis: An Overview of Heart Disease Prediction. *International Journal of Computer Applications*. 2011; Vol.17(8): 43–48. DOI: https://doi.org/10.5120/2237-2860.

4. Ruvinskaya, V. M., Shevchuk, I. & Michaluk, N. "Models based on conformal predictors for diagnostic systems in medicine". *Applied Aspects of Information Technology. Publ. Science and Technical*. Odessa: Ukraine. 2019; Vol. 2(2): 127–137. DOI: https://doi.org/10.15276/aait.02.2019.4.

5. Qin, S. J. "Survey on Data-Driven Industrial Process Monitoring and Diagnosis". *Annual Reviews in Control*. 2012; Vol. 36(2): 220–234. DOI: https://doi.org/10.1016/j.arcontrol.2012.09.004.

6. Liu, X., Ma, L. & Mathew, J. "Machinery Fault Diagnosis Based on Fuzzy Measure and Fuzzy Integral Data Fusion Techniques". *Mechanical Systems and Signal Processing*. 2009; Vol. 23(3): 690–700. DOI: https://doi.org/10.1016/j.ymssp.2008.07.012.

7. MacGregor, J. & Cinar, A. "Monitoring, Fault Diagnosis, Fault-Tolerant Control and Optimization: Data Driven Methods". *Computers & Chemical Engineering*. 2012; Vol.47: 111–120. DOI: https://doi.org/10.1016/j.compchemeng.2012.06.017.

8. Dai, X. & Gao, Z. "From Model, Signal to Knowledge: A Data-Driven Perspective of Fault Detection and Diagnosis". *IEEE Transactions on Industrial Informatics*. 2013; Vol. 9(4): 2226–2238. DOI: https://doi.org/10.1109/TII.2013.2243743.

9. Wen, L., Li, X., Gao, L. & Zhang, Y. "A New Convolutional Neural Network-Based Data-Driven Fault Diagnosis Method". *IEEE Transactions on Industrial Electronics.* 2018; Vol. 65(7): 5990–5998. DOI: https://doi.org/10.1109/TIE.2017.2774777.

10. Saric-Grgic, I., Grubisic, A., Seric, L. & Robinson, T. J. "Student Clustering Based on Learning Behavior Data in the Intelligent Tutoring System". *International Journal of Distance Education Technology.* 2020; Vol. 18(2): 73–89. DOI: https://doi.org/10.4018/IJDET.2020040105.

11. Wong, B. T. & Li, K. C. "A Review of Learning Analytics Intervention in Higher Education (2011-2018)". *Journal of Computers in Education.* 2020; Vol. 7(1): 7–28. DOI: https://doi.org/10.1007/s40692-019-00143-7.

12. Marbouti, F., Diefes-Dux, H. A. & Madhavan, K. "Models for early prediction of at-risk students in a course using standards-based grading". *Computers & Education.* 2016; Vol.103: 1–15. DOI: https://doi.org/10.1016/j.compedu.2016.09.005.

13. Gasevic, D., Dawson, S., Rogers, T. & Gasevic, D. "Learning analytics should not promote one size fits all: The effects of instructional conditions in predicating academic success". *Internet and Higher Education.* 2016; Vol.28: 68–84. DOI: https://doi.org/10.1016/j.iheduc.2015.10.002.

14. He, L., Agard, B. & Trépanier, M. "A classification of public transit users with smart card data based on time series distance metrics and a hierarchical clustering method". *Transportmetrica A: Transport Science.* 2020; Vol. 16(1): 56–75. DOI: https://doi.org/10.1080/23249935.2018.1479722.

15. Tretyakov, D. "A Self-Learning Diagnosis Algorithm Based on Data Clustering". *Intelligent Control and Automation.* 2016; Vol. 07(03): 84–92. DOI: https://doi.org/10.4236/ica.2016.73009.

16. Omran, M., Engelbrecht, A. & Salman, A. A. "An overview of clustering methods". *Intelligent Data Analysis.* 2007; Vol. 11(6): 583–605. DOI: https://doi.org/10.3233/IDA-2007-11602.

17. Friedman, J. H. & Meulmany, J. J. “Clustering Objects on Subsets of Attributes”. *Royal Statistical Society*. 2004; Vol. 66(4): 815–849. DOI: https://doi.org/10.1111/j.1467-9868.2004.02059.x.

18. Amorim, R. C. “Feature Relevance in Ward’s Hierarchical Clustering Using the L_p Norm”. *Journal of Classification*. 2015; Vol. 32(1): 46–62. DOI: https://doi.org/10.1007/s00357-015-9167-1.

19. Obry, T., Travé-Massuyès, L. & Subias, A. “DyClee-C: a Clustering Algorithm for Categorical Data Based Diagnosis”. *DX’19 – 30th International Workshop on Principles of Diagnosis*. Klagenfurt: Austria. November 2019. – Available from: https://hal.laas.fr/hal-02383492/file/DyClee_C_DX19_Final.pdf. – Active link – 20.03.2020.

20. Min, E., Guo, X., Liu, Q., Zhang, G., Cui, J. & Long, J. “A survey of clustering with deep learning: From the perspective of network architecture”. *IEEE Access*. 2018; Vol.6: 39501–39514. DOI: https://doi.org/10.1109/ACCESS.2018.2855437.

21. “Clustering Methods for Big Data Analytics”. *Techniques, Toolboxes and Applications*. Cham, Switzerland: *Springer Nature Switzerland AG*. 2019. 187 p.

22. Shibaev, D., Vyuzhuzhanin, V., Rudnichenko, N., Shibaeva, N. & Otradskaya, T. “Data Control in the Dagnostics and Forecasting the State of Complex Technical Systems”. *Herald of Advanced Information Technology*. *Publ. Science and Technical*. Odessa: Ukraine. 2019; Vol.2 No.3: 183–196. DOI: https://doi.org/10.15276/hait.03.2019.2.

23. Gu, J., Jiang, Z., Fan, W., Wu, J. & Chen, J. “Real-Time Passenger Flow Anomaly Detection Considering Typical Time Series Clustered Characteristics at Metro Stations”. *Journal of Transportation Engineering Part A-Systems*. 2020; Vol. 146(4), article 04020015. DOI: https://doi.org/10.1061/JTEPBS.0000333.

24. Fowlkes, E., Gnanadesikan, R. & Kettenring, J. “Variable Selection in Clustering”. *Journal of Classification*. 1988; Vol.5: 205–228.

25. Gnanadesikan, R., Kettenring, J. & Tsao, S. “Weighting and Selection of Variables for Cluster Analysis”. *Journal of Classification*. 1995; Vol.12: 113–136.

26. Dresp-Langley, B., Ekseth, O. K., Fesl, J., Gohshi, S., Kurz, M. & Sehring, H.-W. “Occam's Razor for Big Data? On Detecting Quality in Large Unstructured Datasets”. *Applied Sciences-Basel*. 2019; Vol. 9(15), article 3065. DOI: https://doi.org/10.3390/app9153065.

27. Huang, J. Z. X., Ng, M. K., Rong, H. Q. & Li, Z. C. “Automated Variable Weighting in k-means Type Clustering”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2005; Vol. 27(5): 657–668. DOI: https://doi.org/10.1109/TPAMI.2005.95.

28. Iam-On, N. “Clustering Data with the Presence of Attribute Noise: a Study of Noise Completely at Random and Ensemble of Multiple k-means Clustering”. *International Journal of Machine Learning and Cybernetics*. 2020; Vol.11: 491–509. DOI: https://doi.org/10.1007/s13042-019-00989-4.

29. Shirkhorshidi, A. S., Aghabozorgi, S. & Wah, T. Y. “A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data”. *PLoS ONE*. 2015; Vol. 10(12), article e0144059. DOI: https://doi.org/10.1371/journal.pone.0144059.

30. Arora, J., Khatter, K. & Tushir, M. “Fuzzy c-Means Clustering Strategies: A Review of Distance Measures”. *Software Engineering: Advances in Intelligent Systems and Computing*. 2015; Vol.731: 153–162. DOI: https://doi.org/10.1007/978-981-10-8848-3_15.

31. Tolentino, J. A., Gerardo, B. D. & Medina, R. P. “Enhanced Manhattan-Based Clustering Using Fuzzy C-Means Algorithm”. *Recent Advances in Information and Communication Technology*. 2019; Vol. 769: 126–134. DOI: https://doi.org/10.1007/978-3-319-93692-5_13.

32. Blackburn, S. R., Bomberger, C. & Winkler, P. “The minimum Manhattan Distance and Minimum Jump of Permutations”. *Journal of Combinatorial Theory Series A*. 2019; Vol.161: 364–386. DOI: https://doi.org/10.1016/j.jcta.2018.09.002.

33. Han, H., Mu, J., He, Y.-C. & Jiao, X. “Cosset Partitioning Construction of Systematic Permutation Codes Under the Chebyshev Metric”. *IEEE Transactions on Communications*. 2019; Vol.67(6): 3842–3851. DOI: https://doi.org/10.1109/TCOMM.2019.2900679.

34. Rossi, G. C. & Testa, M. “Euclidean Versus Minkowski Short Distance”. *Physical Review D*. 2018; Vol.98(5), article 054028. DOI: https://doi.org/10.1103/PhysRevD.98.054028.

35. Krisilov, V., Liubchenko, V. & Kavitska, V. “The Methods for Goal-Oriented Estimation of Model Adequacy”. [Metody tseleorientirovannoi otsenki adekvatosti] (in Russian). *Odes’kyi Politechnichnyi Universytet. Pratsi*. 2012; Vol. 2(39):160–184

36. Krisilov, V. A. & Komleva, N. O. “Analysis and Evaluation of Competence of Information Sources in Problems of Intellectual Data Processing”. [Analiz i ocenka kompetentnosti istochnikov informacii v

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

zadachah intellektual'noj obrabotki dannyh] (in Russian). *Problemele Energeticii Regionale*. 2019; Vol.1-1(40): 91–104 DOI: https://doi.org/10.5281/zenodo.3239184

37. Larshin, V. & Lishchenko, N. "Educational technology information support". *Herald of Advanced Information Technology. Publ. Science i Technical*. Odessa: Ukraine. 2019; Vol.2 No.4: 317–327. DOI: https://doi.org/10.15276/hait.04.2019.8.

Conflicts of Interest: the authors declare no conflict of interest

Received 25.11.2019
Received after revision 14.02.2020
Accepted 20.02.2020

DOI: https://doi.org/10.15276/aait.01.2020.1
УДК 004.9

МЕТОДОЛОГІЯ ІНФОРМАЦІЙНОГО МОНІТОРИНГУ ТА ДІАГНОСТИКИ ОБ'ЄКТІВ, ПРЕДСТАВЛЕНИХ КІЛЬКІСНИМИ ОЦІНКАМИ, З ВИКОРИСТАННЯМ КЛАСТЕРНОГО АНАЛІЗУ

Наталія Олегівна, Комлева[1)]
ORCID: http://orcid.org/0000-0001-9627-8530, komleva@opu.ua, Scopus ID: 57191858904
Vira V. Liubchenko[2)]
ORCID: http://orcid.org/0000-0002-4611-7832, lvv@opu.ua. Scopus ID: 56667638800
Світлана Леонідівна Зіноватна[1)]
ORCID: http://orcid.org/0000-0002-9190-6486 zinovatnaya.svetlana@opu.ua, Scopus ID: 57206667710
[1)] Одеський національний політехнічний університет, пр. Шевченка, 1.Одеса, 65044, Україна
[2)] Hochschule für Angewandte Wissenschaften Hamburg, Fakultät Life Sciences, Ulmenliet 20 Hamburg, Germany

АНОТАЦІЯ

В роботі обговорюються методологічні основи інформаційного моніторингу та діагностики з використанням кластерного аналізу для класу об'єктів, опис яких представлений кількісними оцінками. Аналіз публікацій показав, що застосування кластерного аналізу для визначення станів об'єктів в окремих випадках було успішним, до того ж теорія кластерного аналізу добре розроблена, а властивості методів вивчені, що свідчить про доречність використання апарату кластерного аналізу. Отже розробка узагальненої методології для діагностування будь-яких об'єктів, опис яких визначаються вектором оцінок, є актуальною задачею. Метою роботи є розробка методологічних основ визначення діагностичних станів та поведінкових шаблонів для об'єктів, які описано кількісними ознаками, за допомогою кластерного аналізу. Оскільки інформаційна діагностика – це цілеспрямована діяльність по оцінці стану об'єкту дослідження на основі динамічної інформаційної моделі, спочатку обговорюється модель об'єкту діагностування. При цьому розглядається життєвий цикл об'єкту діагностування, що описуються множиною параметрів, значення яких визначаються часовим зрізом на лінії життя екземпляру. Показано, що кожний стан об'єкту діагностування може характеризуватися різною кількістю значень мір. Для запобігання передчасного переривання життєвого циклу екземпляру визначені характеристики, аналіз яких дозволяє діагностувати такий стан екземпляру або траєкторію його поведінки, що може свідчити про загрозу існування екземпляру та необхідність прийняття підтримуючих процедур. Формалізація умов для припинення життєвого циклу досліджуваного об'єкта та формування переліку підтримуючих процедур здійснюється експертним чином. Якість будь-якої інформаційної технології залежить від якості вхідних даних, тому розроблено процедуру аналізу діагностичних ознак для визначення адекватності моделі об'єкту діагностування. Розроблені методології одно-, двох- та N-крокового діагностування на базі значень центрів кластерів, що дає можливість розпочинати діагностування якомога раніше та застосовувати доступні дані якомога повніше. В усіх процедурах використано відношення порядку на кластерах. Для процедури двокрокового діагностування визначено паттерни переходів, що дозволяють визначати зміни станів об'єкту діагностування, для N-крокового – паттерни трендів. Процедура аналізу діагностичних ознак та визначення зазначених паттернів є новими науковими результатами. Застосування розроблених процедур показано на прикладі діагностування успішності студентів. При цьому у якості моделі предметної області застосовано освітню програму за обраної спеціальності. Для однокрокової діагностики досліджено наявність впливу латентного фактору та діагностичних ознак, що демонструють суттєву нестійкість. Для одно- та двокрокової методології надано умови формування сегменту ризику.

Ключові слова: інформаційна діагностика; кластерний аналіз; діагностична ознака; паттерн, тренд

DOI: https://doi.org/10.15276/aait.01.2020.1
УДК 004.9

МЕТОДОЛОГИЯ ИНФОРМАЦИОННОГО МОНИТОРИНГА И ДИАГНОСТИКИ ОБЪЕКТОВ, ПРЕДСТАВЛЕННЫХ КОЛИЧЕСТВЕННЫМИ ОЦЕНКАМИ, С ИСПОЛЬЗОВАНИЕМ КЛАСТЕРНОГО АНАЛИЗА

Наталия Олеговна Комлевая[1)]
ORCID: http://orcid.org/0000-0001-9627-8530, komleva@opu.ua. Scopus ID: 57191858904

Vira V. Liubchenko[2)]
ORCID: http://orcid.org/0000-0002-4611-7832, lvv@opu.ua. Scopus ID: 56667638800
Светлана Леонидовна Зиноватная[1)]
ORCID: http://orcid.org/0000-0002-9190-6486, zinovatnaya.svetlana@opu.ua. Scopus ID: 57206667710
[1)] Одесский национальный политехнический университет, пр. Шевченко, 1. Одесса, 65044, Украина
[2)] Hochschule für Angewandte Wissenschaften Hamburg, Fakultät Life Sciences, Ulmenliet 20 Hamburg, Germany

АННОТАЦИЯ

В работе обсуждаются методологические основы информационного мониторинга и диагностики с использованием кластерного анализа для класса объектов, описание которых представлено количественными оценками. Анализ публикаций показал, что применение кластерного анализа в отдельных случаях было успешным, к тому же теория кластерного анализа хорошо разработана. Поэтому разработка обобщенной методологии для диагностики объектов, описание которых представляется вектором количественных признаков, является актуальной задачей. Поскольку информационная диагностика – это целенаправленная деятельность по оценке состояния объекта исследования на основе динамической информационной модели, первоначально формализована модель объекта диагностирования при помощи множества параметров, значения которых определяются временными срезами на линиях жизни для каждого экземпляра объекта диагностирования. Разработаны методологии одно-, двух и N-шагового диагностирования на базе значений центроидов кластеров, что дает возможность наиболее полно использовать имеющиеся данные и оценивать статику состояний и динамику изменений этих состояний. Применение разработанных методологий показано на примере диагностики успешности студентов.

Ключевые слова: информационная диагностика; кластерный анализ; диагностический признак; паттерн; тренд

ABOUT THE AUTHORS

Nataliia O. Komleva, PhD (Eng), Associate Professor of System Software Department, Odesa National Polytechnic University, 1, Shevchenko, Avenue. Odessa, 65044, Ukraine
komleva@opu.ua. Scopus ID: 57191858904. ORCID: http://orcid.org/0000-0001-9627-8530
***Research field*:** Data Analysis, Software Engineering, Knowledge Management

Наталія Олегівна Комлева, кандидат техніч. наук, доцент каф. Системного програмного забезпечення. Одеський національний політехнічний університет, пр. Шевченка, 1. Одеса, 65044, Україна

Vira V. Liubchenko – Doctor of Engineering Sciences, Professor; Lecturer, Fakultät Life Sciences, Hochschule für Angewandte Wissenschaften Hamburg, Ulmenliet 20. Hamburg, 21033, Germany
ORCID: https://orcid.org/0000-0002-4611-7832; lvv@op.edu.ua. Scopus Author ID: 56667638800
Research field: Data Science; Software Engineering; Project Management

Svitlana L. Zinovatna, PhD (Eng), Associate Professor of System Software Department, Odesa National Polytechnic University, 1, Shevchenko, Avenue. Odesa, 65044, Ukraine
zinovatnaya.svetlana@opu.ua. Scopus ID: 57206667710 .ORCID: http://orcid.org/0000-0002-9190-6486
***Research field*:** Data Analysis, Information System Productivity

Світлана Леонідівна Зіноватна, кандидат техніч. наук, доцент каф. Системного програмного забезпечення. Одеський національний політехнічний університет, пр. Шевченка, 1.Одеса, 65044, Україна

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)