Two Step Density-Based Object-Inductive Clustering Algorithm

Volodymyr Lytvynenko[1][0000-0002-1536-5542], Irina Lurie[1][0000-0001-8915-728X], Jan Krejci[2][0000-0003-4365-5413], Mariia Voronenko[1][0000-0002-5392-5125], Nataliia Savina[3][0000-0001-8339-1219], Mohamed Ali Taif [1][0000-0002-3449-6791]

[1]Kherson National Technical Uneversity, Kherson, Ukraine,
[2]Jan Evangelista Purkyne University in Usti nad Labem, Czech Republic,
[3]National University of Water and Environmental Engineering,
Rivne, Ukraine,
immun56@gmail.com,lurieira@gmail.com,mary_voronenko@i.ua,
taifmohamedali@gmail.com, jan.krejci@ujep.cz,
n.b.savina@nuwm.edu.ua

Abstract. The article includes the results of study into the practical implementation of two-step DBSCAN and OPTICS clustering algorithms in the field of objective clustering of inductive technologies. The architecture of the objective clustering technology was developed founded on the two-step clustering algorithm DBSCAN and OPTICS. The accomplishment of the technology includes the simultaneous data's clustering on two subsets of the same power by the DBSCAN algorithm, which involve the same number of pairwise objects similar to each other with the subsequent correction of the received clusters by the OPTICS algorithm. The finding the algorithm's optimal parameters was carried out based on the clustering quality criterion's maximum value of a complex balance, which is rated as the geometric average of the Harrington desirability indices for clustering quality criteria (internal and external).

Keywords: Clustering, Density-based clustering, Objective clustering, Inductive clustering, clustering quality criteria, Two Step Clustering, DBSCAN, OPTICUS

1 Introduction

Clustering is the primary method for extracting data. The task of clustering is a special case of the task of learning without a teacher and reduces to break the set of data objects into subsets so that the elements of one subset are significantly different in some set of properties from the elements of all other subsets. Clustering can be a preprocessing step in other data extraction applications.

There are many different clustering algorithms. Some of them divide the set into a known amount of clusters, but some of them automatically select the amount of clusters.

The density-based algorithm is a highly efficient and simple algorithm [1]. Different methods are best suited for different databases. In this paper, we consider the DBSCAN and OPTICS clustering algorithms, which are used to find clusters of various shapes, densities and sizes in spatial data sets with noise.

The clustering algorithm, Named DBSCAN, (Density Based Spatial Clustering of Applications with Noise) was proposed in [1]. It is based on the assumption that the density of points, which are located inside the clusters, is greater than behind the clusters. This algorithm allows finding nonlinearly separable clusters of arbitrary shape. It can detect clusters completely encircled, but not connected with other clusters. It does not need specification of the amount of clusters, distinguishes noise and is resistant to outliers.

However, the DBSCAN algorithm is not without flaws. The boundary points that can be reached from more than one cluster can belong to any of these clusters, which rely on the order of viewing the points.

OPTICS clustering algorithm (Ordering points to identify the clustering structure) as well as DBSCAN allows finding clusters in data space based on density and was proposed in [2]. However, unlike DBSCN, this algorithm uses the distance between neighboring objects to obtain the availability field, which is used to separate clusters of different densities from noise, which solves the problem of finding content clusters in data that have different densities. To do this, the data is ordered, so that the spatially close points become adjacent in the ordering. For each point, a special distance is stored that represents the density that should be taken for the cluster so that the points belong to the same cluster. The result of this procedure is presented in the form of a dendrogram.

Algorithms based on density are highly efficient and simple algorithms [1]. Different methods are best suited for different databases. Here we are dealing with DBSCAN and OPTICS, which are used to find clusters of various shapes, densities and sizes in spatial data sets with noise.

The idea underlying this algorithm is that inside each cluster there is a typical density of points (objects), which is noticeably higher than the density outside the cluster, as well as the density in areas with noise lower than the density of each cluster.

On the other hand, inductive clustering methods [3] allow for inaccurate noisy data and short samples, using the minimal amount of the chosen quadratic criterion, to find a non-physical model (decision rule), the accuracy of which is less than the structure of the full physical model.

Examining the set of candidate models by external criteria is necessary only for non-physical models. In case of small dispersion of interference, it is advisable to use internal search criteria. With increasing interference, it is advisable to move to nonparametric algorithms. The use of inductive clustering methods is advisable because they almost always ensure that the optimal amount of clusters is found that is adequate for the noise level in the data sample.

The main idea of this work is to combine the density algorithms DBSCAN and OPTICS, which allow you to recognize clusters of various shapes, as well as define content clusters for data with different densities and in the form of a two-step algorithm and an inductive clustering method that will significantly improve the accuracy when recognition of complex objects. It is assumed that by combining these methods, it is possible to solve some of the problems listed above with a sufficiently high result.

The aim of the work is to develop a methodological basis for constructing hybrid inductive cluster-analysis algorithms for isolating (clustering) objects with complex non-linear forms with high recognition accuracy and resolution.

2 Review of the Literature

The classification of several clustering algorithms by their categories is presented in [4]. Each of them has its advantages and disadvantages. The choice of an appropriate clustering algorithm is definited by the type of data being examined and the purpose of the current task.

Non-parametric algorithms capable of distinguishing clusters of arbitrary shape also allow obtaining a hierarchical representation of data.

The approach used by these algorithms for non-parametric density estimation is that the density is characterized by the number of nearby elements [5]. Thus, the proximity of a pair of elements is determined by the amount of common neighboring elements. The most prominent representative of this approach is the DBSCAN clustering algorithm [6].

Its basic idea is that if an element in the radius keeps a specified amount (*MinPts*) of neighboring elements, then all its “neighbors” are placed in the same cluster with it. Elements that do not have a sufficient number of "neighbors" and are not included in any cluster belong to "noise". DBSCAN allows you to select clusters of complex shape and cope with the choices and "noise" in the data. The disadvantages of the algorithm are the complexity of setting parameter values (and *MinPts*) [7] and the difficulty in identifying clusters with significantly different densities.

The OPTICS algorithm [8] is a generalization of DBSCAN, where elements are ordered into a spanning tree so that the spatially close elements are close together. In this case, there is no need to carefully adjust the appropriate parameter, and the result is a hierarchical result [5]. One of the major drawbacks of the existing clustering algorithms is the reproducibility error. The basic idea for solving this problem was proposed in [9].

In [10,11], the authors showed that a decrease in reproducibility error can be achieved through the use of inductive modeling methods for complicated systems, which are a logical prolongation of group data processing methods. The issues of creating a methodology for analyzing inductive systems as a tool for analytical planning of engineering research are considered in [12]. In [13], the authors first proposed a hybrid inductive clustering algorithm based on DBSCAN.

The work [14] presents the results of computational experiments using objective cluster inductive technology of multidimensional high-dimensional data. The authors showed that the implementation of this technology based on some clustering algorithm involves determining the affinity function between objects, clusters, and objects, and clusters at the first stage. Then we need to share the investigated data into two subsets of the same power, which contain the same number of pairs of similar objects. The formation of quality criteria for the clustering of internal, external and complex balance should be carried out at the next stage. Optimal clustering is determined on the basis of the extreme values of the criteria used in the sequential enumeration of admissible clustering.

The article [15] describes the study's results of the practical accomplishment of the DBSCAN clustering algorithm within the objective clustering of inductive technology. In this paper, the finding of the optimal parameters of the algorithm was performed by use of the complex criterion maximum value for the quality of clustering, which is calculated as the geometric average of the indicators of the desirability of Harrington for external and internal criteria for the quality of clustering.

The work [16] investigated the problem of clustering complex data of inductive objective clustering technology. A practical accomplishment of the hybrid data clustering model based on the integrated use of R and KNIME software tools has been implemented. The model performance was evaluated using various types of data. The simulation results showed the high efficiency of the proposed technology. It is shown that the proposed method allows reducing the reproducibility error value because the final decision on determining the optimal parameters of the clustering algorithm is made on the basis of parallel analysis of clustering results obtained on equally powerful data sets taking into account the difference in clustering results obtained on these subsets.

In this article, we describe a hybrid model of an objective cluster inductive technology founded on the two-step clustering algorithm DBSCAN and OPTICS. The practical implementation of the proposed model was performed on R.

3 Problem Statement

The formulation of the clustering problem is as follows: let X is the set of objects, Y is the set of amounts (names, labels) of clusters. The function of the distance between objects $\rho(x,x')$ is also set. It is necessary to divide the sample into subsets that do not overlap (clusters), so that each cluster composes of objects ρ that are close in metric, and the objects of different clusters are significantly different. In addition, each object $x_i \in X^m$ corresponds to a cluster number y_i. In this case, the clustering algorithm can be considered as a function $a: X \to Y$ that assigns a cluster number $y \in Y$ to any object $x \in Y$. In some cases, the set Y is known in advance, but more often the task is to find the optimal amount of clusters according to one or another criterion of the quality of clustering [3].

In inductive clustering methods, the cluster model is selected using the minimum external balance criterion, which characterizes the quality of clustering of the corresponding model on two identical power sets.

Formally, the optimal inductive clustering model can be presented as:

$$M : \left\{ R(K) \mid e \le e_0, \tau \le \tau_0 \xrightarrow{\{CR\}} opt \right\} \tag{1}$$

$R(K)$ is the result of clustering, e is the error of clustering on the training and test samples, τ is the time interval of the clustering process, *CR* is the set of internal and external criteria for assessing the quality of clustering.

4 Materials and Methods

4.1 DBSCAN Clustering Algorithm

The idea underlying the algorithm is that inside each cluster there is a typical density of points (objects), which is noticeably higher than the density outside the cluster, as well as the density in areas with noise below the density of any of the clusters. For each point of the cluster, its neighborhood of a given radius must contain at least a certain amount of points, this amount of points is specified by a threshold value.

Most algorithms that produce a flat partition create clusters in the form close to spherical, since they minimize the interval of documents to the center of the cluster [17].

DBSCAN authors have shown experimentally that their algorithm is capable of recognizing clusters of different shapes. The basic idea behind the algorithm lies in the fact that within each cluster there is a typical density of points (objects) that is noticeably higher than the outside density of the cluster, as well as the density in areas with noise below the density of any of the clusters. Even more precisely, for each point of the cluster, its neighborhood of a given radius must contain some amount of points, this amount of points is given by the limit values [18].

The basis of this algorithm is several definitions [18]:

- ε is the vicinity of the object is called the outskirts of the radius ε of some object;
- the root object is named an object ε is the neighborhood of which contains some minimum amount of *MinPts* objects;
- the object p is directly tightly accessible from the object q if p located in ε is the neighborhood q and q is the root object;
- the object p is the tight reachable from the object q for the given ε and the parameter *MinPts*, if there is a sequence of objects $p, \dots, p$, where $p = q$ and $p = p$ such that $p+1$ is directly densely achievable with p, $1 \le i \le n$;
- the object p is tightly connected to the object q when given ε and *MinPts*, if there is an object o that p is the same as q the available volume from o.

To search for clusters, the DBSCAN algorithm checks ε is the neighborhood of each object. If ε is the neighborhood of the object p contains more points than *MinPts*, then a new cluster with a root object p created. Then, DBSCAN iterative collects objects directly tightly reachable from the root objects, which can lead to the union of several tight reachable clusters. The process is completed when no new object can be added to one cluster.

Although the DBSCAN algorithm does not need the pre-specified amount of clusters received, it will be necessary to specify parameters values ε and *MinPts* that directly affect the clustering result. The optimal values of these parameters are difficult to determine, especially for multidimensional data spaces. In addition, the distribution of data in such spaces is often asymmetric, which does not allow them to be used for clustering of global density parameters.

The work of the DBSCAN algorithm is as follows.

```
Enter: the set of objects S, Eps and MinPt.
An object can be in one of three states:
    1. Not noted.
    2. It is noted that no cluster is the internal object.
    3. Attributed to some cluster.
Step 1. To set all the elements of the set S flag S "not marked". Assign the current
        cluster $C_j$ to a zero number, $j = 0$. The set of noise points Noise = 0.
Step 2. For each $s_i \in S$ such flag $(s_i)$ = "not marked", execute:
Step 3. Flag $(s_i)$ = "not marked";
Step 4 $N_i = N_{Eps}(s_i) = \{q \in S | dist(s_i, q) \leq Eps\}$
Step 5. If $|s_i| < MinPt$, then $Noise = Noise + \{s_i\}$
        Otherwise the number of the next cluster $j = j + 1$;
EXPANDCLUSTER $(s_i, N_i, C_j, Eps, MinPt)$;
Exit: The set of clusters $C = (C_j)$.
EXPANDCLUSTER
Login: The current object $s_i$, its eps neighbor $N_i$, the current cluster $N_i$ and
        $Eps, MinPt$.
Step 1 $C_j = C_j + \{s_i\}$;
Step 2. For all points $s_k \in N_i$:
Step 3. If the flag $(s_k)$ = "not marked", then
Step 4 flag $(s_k)$ = "marked";
Step 5. $N_{ik} = N_{Eps}(s_k)$;
Step 6. If $|N_{ik}| \geq MinPt$, then $N_i = N_i + N_{ik}$;
Step 7. If $\nexists\ p : s_k \in C_p, p = \overline{1,(C)}$, those $C_j = C_j + \{s_k\}$;
Exit: cluster $C_j$.
```

As the research shows [18], the considered clustering algorithm has a number of advantages that make it possible to use this method for working with clusters of different nature (forms); the application of this algorithm allows you to work with large-scale samples and allows you to work with n-dimensional objects (these are objects whose attributes are more than 3 if the function is appropriately selected for calculating the distance (in the general case it is possible to use the Markov metric) However, a significant disadvantage is a rather laborious procedure for determining the required parameters for the correct operation of the algorithm. More detailed descriptions and drawbacks of the DBSCAN algorithm are shown in Table 1.

Table 1. Advantages and disadvantages of the DBSCAN algorithm

Advantages	Disadvantages
1. DBSCAN can find arbitrary clusters. 2. DBSCAN has a notion of noise and is resistant to emissions, i.e. all emissions are made in a separate cluster. 3. Does not need a priori task of the amount of clusters, in contrast to the K-mean algorithm. 4. Uses only two parameters and is basically not sensitive to the ordering of points in the database. 5. Allows working with samples of large dimensional data. 6. Defining the parameters *MinPts* and ε allow working with n-dimensional objects provided that an appropriate function is selected for the calculation of the distance.	1. DBSCAN is not a fully deterministic algorithm: the boundary points that can be accessed from more than one cluster may be part of another cluster, depending on the order of data processing. 2. The quality of DBSCAN operation depends on the distance used. The most commonly used Euclidean distance; But for multidimensional data, this indicator can be almost useless due to the so-called "curse of dimension", which makes it difficult to find the nearest value for ε. This effect is also located in any other algorithm based on the Euclidean distance. 3. DBSCAN cannot copy the data to a large difference in density, since the combination *MinPts* ε cannot be selected appropriately for all clusters. 4. If the scale and data are not understandable, it may be very difficult to choose a significant distance from the threshold ε. 5. Significant drawback is a very laborious procedure for determining the required parameters for the correct algorithm procedure.

In the general case, the DBSCAN algorithm has a quadratic computational complexity due to the search for the Eps is neighborhood. However, the authors of the algorithm used a special data structure for this purpose R * are trees, as a result, the search for Eps is neighborhood for one point O (log n). The total computational complexity of DBSCAN is O (n * log n) [19].

4.2 OPTICS Clustering Algorithm

The concept of the OPTICS algorithm [8] is similar to DBSCAN, but the algorithm is designed to get rid of one of the main weaknesses of the DBSCAN algorithm is the problem of finding content clusters in data that has different densities.

To do this, the database points are (linearly) ordered so that the spatially close points become adjacent in the ordering. In addition, for each point, a special distance is stored that represents the density that should be taken for the cluster, so that the points belong to the same cluster. This is presented in the form of a dendrogram.

In this case, there is no need to carefully adjust the appropriate parameter, and the result is a hierarchical result [5]. However, the parameter is specified in the algorithm as the maximum radius considered. Ideally, it can be set very large, but this leads to exorbitant computational costs.

OPTICS density algorithm [8] also allows you to select a hierarchical structure and clusters of complex shape. The data is ordered into a spanning tree so that the spatially close elements are located nearby. In this case, the hierarchy is represented in the form of a reachability diagram, on which the reachability distances for the constructed sequence of elements are marked. The peaks in the diagram correspond to the divisions between the clusters, and their height to the distance. If necessary, a dendrogram can be easily constructed from a reachability diagram. Since for each element only adjacent elements in a limited radius ε are considered, the OPTICS algorithm can be implemented with computational complexity, which is not sufficient for processing large data arrays.

DBSCAN requires two parameters, the optimal values of which are difficult to determine. Therefore, an OPTICS algorithm was proposed in [8], which makes it possible to order the initial set and simplify the clustering process. In accordance with it, a reachability diagram is constructed, thanks to which, with a fixed *MinPts* value, it is possible to process not only the specified value e, but also all e * <e.

Unlike DBSCAN, the OPTICS algorithm also considers points that are also part of a denser cluster, so each point is assigned a main distance, which describes a distance, which describes the distance to the *MinPts* -th nearest point:

$$core-dist_{\varepsilon,MinPts}=\begin{cases}UNDEFINED & \left|N_{\varepsilon}\left(p\right)\right|<MinPts\\ MinPts\text{–}thN_{\varepsilon}\left(p\right) & \left|N_{\varepsilon}\left(p\right)\right|\geq MinPts\end{cases}\qquad(2)$$

$core-dist_{\varepsilon,MinPts}$ is the main interval and $MinPts\text{–}thN_{\varepsilon}\left(p\right)$ is the ascending order of interval to $N_{\varepsilon}\left(p\right)$.

The attainable interval of a point *o* from a point *p* is equal to either the interval between *p* and *o*, or the main interval of the point *p*, depending on which value is greater:

$$reachability-dist_{\varepsilon,MinPts}\left(o,p\right)=\begin{cases}UNDEFINED & \left|N_{\varepsilon}\left(p\right)\right|<MinPts\\ \max\left(core-dist_{\varepsilon,MinPts(p)},dist\left(p,o\right)\right) & \left|N_{\varepsilon}\left(p\right)\right|\geq MinPts\end{cases}\qquad(3)$$

$reachability-dist_{\varepsilon,MinPts}\left(o,p\right)$ is the attainable interval. If *p* and *o* are the nearest neighbors, and $\varepsilon'<\varepsilon$, we can assume that *p* and *o* belong to the same cluster.

Both the main and achievable interval s are not determined if there is not a sufficiently dense cluster (applied to ε). If you take a large enough, this will never happen, but then any query ε is the neighborhood returns the entire database, which leads to work time $O\left(n^2\right)$. The parameter ε is required to cut loose clusters that are no longer interesting, and thereby speed up the algorithm. The parameter ε, strictly speaking, is optional. It may simply be set to the maximum possible value. However, when a spatial index is available, it affects the computational complexity. OPTICS differs from DBSCAN in that this parameter is not taken into account, if it can influence, then only in that it sets the maximum value.

The advantage of the algorithm is that it can efficiently process clusters if the data has different densities and retrieves objects in a specific order using the ordering mechanism. The disadvantages of the algorithm include the fact that it is less sensitive to erroneous data than DBSCAN.

4.3 Inductive Clustering Algorithm

Among the main principles of inductive modeling of complex systems are the three mentioned above, namely [20]: the principle of self-organization; the principle of external complement and the principle of freedom of decision-making.

The principle of self-organization of models based on the inductive approach to simulation of complex systems, the origins of which are presented in this topic, categorically rejects the path of expansion and complication of the model and increase the output volume of information about the object and postulates the existence of an optimal, scaled modeling area, and also one model of optimal complexity. It can be synthesized with the help of self-organization, that is, the search for many model applicants for appropriately selected external selection criteria for models. Optimization of the model for some ensemble of criteria determines the results of simulation at the given levels of noise and volume of observations.

The principle of external complement is connected with Godel's theorem "... only external criteria, based on new information, allow us to synthesize the true model of the object, hidden in the data that is noisy." In other words, we can say that, according to this principle, only external criteria (i.e., calculated on the basis of "fresh" data not used for the synthesis of the model) with increasing complexity of the model pass through the minima. The application of this principle is realized by dividing the original data table into two parts A and B.

The principle of freedom of decision-making. In accordance with this principle, for each generation (or series of model selection) there is a certain minimum of selected combinations, which are called freedom of choice and ensure the convergence of multi-row selection of the optimal complexity model. The principles of the freedom to make decisions and the step-by-step (multi-faceted) decision-making procedure are first implemented in the perceptron. The perceptron consists of several customizable link lines. After each series of links, a special device is required that passes the most probable solutions in the next series. On the last placement a single and final decision is taken. In other words, following the purposeful selection of models to determine the optimal complexity model in accordance with the principles set out, the following rules must be observed:

- for each generation (or series of selection) models there is a certain minimum of selected combinations, which are called freedom of choice;
- too many generations leads to an induction (the information matrix becomes poorly defined);
- the more difficult the problem of selection, the more generations need to obtain a model of optimal complexity;
- the freedom of choice is ensured by the fact that for each subsequent series of selection not only one solution is passed, but a few of the best, selected in the last row D.

Gabor formulated this principle in the following way: to make decisions at the given time is necessary in such a way that at the next moment of time when the need for the next decision will arise, freedom of decision-making would be preserved [21].

These principles formed the basis of technology for solving the problems of inductive synthesis of models according to experimental data. The most general formulation of the problem of inductive synthesis of models by experimental data, or structural-parametric identification, is given in [22]. According to these papers, such a statement is to find the extremum of some criterion on the set of different models $\Im$:

$$f^{*} = \arg\min CR(f) \tag{4}$$

Since (1) is not completed by the formulation of the problem, it needs to be further identified, in particular:

- ask a priori expert or expert information about the type, character, and volume of the initial information to be known from the analysis of the experiment;
- specify the class of basic functions from which the set $\Im$ must be formed;
- determine the method of generating models f;
- specify a method for evaluating parameters;
- specify a model comparison criterion $CR(f)$ and specify a method for minimizing it.

In [9], it is noted that the view on clustering as a model allows us to transfer to the theory of cluster analysis all the basic concepts of the theory of self-organization of models based on the method of the group method of data handling (GMDH). Self-organization of clustering models is called their selection in order to choose optimal clustering. The more inaccurate data is the easier it is to optimize clustering (complexity is measured by the amount of clusters and the amount of attributes). In cluster analysis (OCA) algorithms, clusters are formed by the internal criterion (the more complex, the more precise), and their optimal number and composition of the ensemble of attributes are calculated by the external criterion (forming a minimum in the region of under-complicated clusterization, optimal for a given level of dispersion of noise). The overlapping of clustering variants implements the OCA algorithm [23]. The construction of a hierarchical tree of clustering organizes and reduces busting, and the optimal clustering criterion is not lost. Physical clustering is based on the criterion of clustering balance. To calculate the criterion, the sample of data is parted into two equal parts. Each clustering tree is constructed on each sub-sample and the balance criterion is calculated at each step with the same number of clusters. The criterion needs a clusterization in which the number and coordinates of the centers (midpoints) corresponding to each other clusters will coincide [24]:

$$BL = \frac{1}{MK} \sum_{j=1}^{M} \sum_{i=1}^{K} (x_{oA} - x_{oB})^2 \to \min \tag{5}$$

K is the number of clusters in this step of constructing a tree; *M* is the coordinate number; x_{oA} are the coordinates of cluster centers constructed on part *A*; x_{oB} are the coordinates of cluster centers constructed on part *B*.

Table 2. Advantages and disadvantages of inductive clustering algorithm

Advantages	Disadvantages
1. The optimal complexity of the structure of the model is found, which is adequate to the level of obstacles in the data sample. 2. Guaranteed finding is the most accurate or non-contained model. The method does not miss the best decision when checking all options. 3. Does not require the task of the model in explicit form, the model is constructed itself, in the process of the algorithm. 4. The method works on short samples, when the number of coefficients of the model is less than the points of observation (A small amount of empirical information). 5. High accuracy of forecast. 6. Minimizing the influence of subjective factors when constructing the model. 7. Low cost model. 8. Realization of the logic of "discovery". 9. Ability to adjust the forecast for new facts. 10. Construction of an objective model during the operation of the algorithm. 11. High impedance: algorithms do not lose performance at a signal / noise ratio at θ = 300-400%.	1. At close experimental points, the phenomenon of degeneration of the matrix of normal Gaussian equations is possible, as a result of which there is a need for the use of special methods of regularization. 2. Complete absence of explanatory function. 3. Low quality of intuitive user formation. 4. The impossibility of constructing a model for random and pseudorandom behavior of objects. 5. Eurasticity of some self-organization procedures. 6. The main limitation in the work of the method of group accounting of arguments is the large volume of selected options, resulting in a slow convergence of the method and a significant length of time of its work. 7. Relatively high computing value. 8. Since instead of a complete search it uses truncation, you cannot find the right model.

To determine the quality of clustering, both internal criteria and an external balance criterion were used.

The internal quality criteria for clustering were:

1. Index Dunn [25] compare the cluster spacing with the cluster's diameter. The higher the index value, the better the clustering.

$$DI\left(K\right)=\min_{i\in K} \qquad (6)$$

2. Index Calinski – Harabasz [26]

$$QC_{CH}=\frac{QCB\cdot\left(N-K\right)}{QCW\cdot\left(K-1\right)}\to\max \qquad (7)$$

N is the amount of objects, *K* is the amount of clusters. The maximum value of the index corresponds to the optimal cluster structure.

For the calculation of the external criterion of balance, the approach taken in [27] was taken as the basis for the basis. In this paper, the external criterion of the (*EC*) controlled clusterization is defined as the normalized optimal value of the sum of the squares of deviations between the values of the internal criteria (*IC*) of the clustering quality (1) - (2):

$$ECB\Big|_{K_A=K_B}=\sqrt{\frac{\left(IC_A-IC_B\right)^2}{\left(IC_A+IC_B\right)^2}} \qquad (8)$$

To create equal conditions for clustering on subsets and when using the DBSCAN clustering algorithm, an equal amount of clusters is determined at the clustering stage. The module of the difference in the values of the external balance criteria with the same amount of clusters on each subset reaches a minimum value:

$$\left|ECB_{K_P}-ECB_{K_{P+1}}\right|\to\min \qquad (9)$$

K_P and K_{P+1} are the number of clusters *P* and *P*+1. For each K_P and K_{P+1}, it is fixed eps_{K_P} и $eps_{K_{P+1}}$ to define clusters on the set Ω :

$$\begin{aligned}&eps\in\left[eps_{K_P},eps_{K_{P+1}}\right],\Delta eps=0,001\\&minPts\in\left[minPts_{\min},minPts_{\max}\right],\Delta minPts=1\end{aligned} \qquad (10)$$

To get rid of one of the main weaknesses of the DBSCAN algorithm is the problem of finding content clusters in data that have different densities, we will further use the OPTICS algorithm.

4.4 Two Step Density-Based Objective Inductive Technology Based on DBSCAN and OPTICS Clustering Algorithm

The main idea of this study is the combined use of a hybrid architecture that combines several computational paradigms, the main focus of which is on obtaining synergistic effects from their combination or, in other words, hybridization. In a hybrid architecture that combines several paradigms, the effectiveness of one approach can compensate for the weakness of the other (Fig. 1).

By combining different approaches, it is possible to circumvent the disadvantages inherent in each separately. Hybrid algorithms usually consist of various components that are combined in the interests of achieving their goals.

In our study, data processing begins with dividing the studied data into two equally powerful subsets using inductive objective clustering based on DBSCAN, then the definition of meaningful clusters is performed using the OPTICS algorithm for data with different densities (Fig. 1)

The integration and hybridization of various methods and information technologies makes it possible to solve complex problems that cannot be solved on the basis of any particular methods or technologies. In this case, in the case of the integration of heterogeneous information technologies, one should expect synergistic effects of a higher order than when combining various models within one technology.

Hybridization helps to take advantage of each of the interacting components while reducing the effects of their disadvantages and limitations. Hybrid intelligent systems, that is, those that combine several components, have recently attracted considerable attention due to their ability to solve complex problems that are characterized by inaccuracies, uncertainty, unpredictability, high dimensionality and environmental variability. They can use both expert knowledge and raw data, often providing original and promising ways to solve problems.

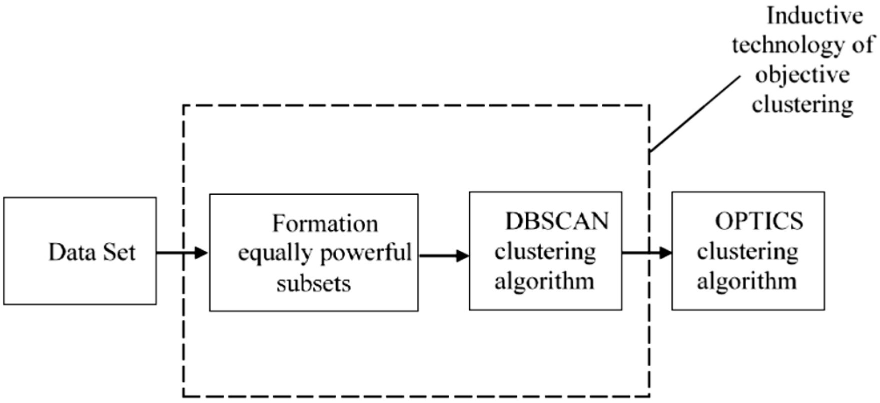
Inductive technology of objective clustering
Data Set
Formation equally powerful subsets
DBSCAN clustering algorithm
OPTICS clustering algorithm

Fig. 1. Generalized scheme of the proposed hybrid two-step inductive clustering algorithm based on DBSCAN and OPTICS.

The more detailed diagram of the proposed hybrid objective clustering technology is shown in Fig. 2. This includes such steps:

Step 1. Start

Step 2. Formation of the initial set of objects Ω under study. Representation of data in the form of an n × m matrix, where n is the amount of rows or the amount of objects studied, is the amount of columns or the amount of features that characterize the objects.

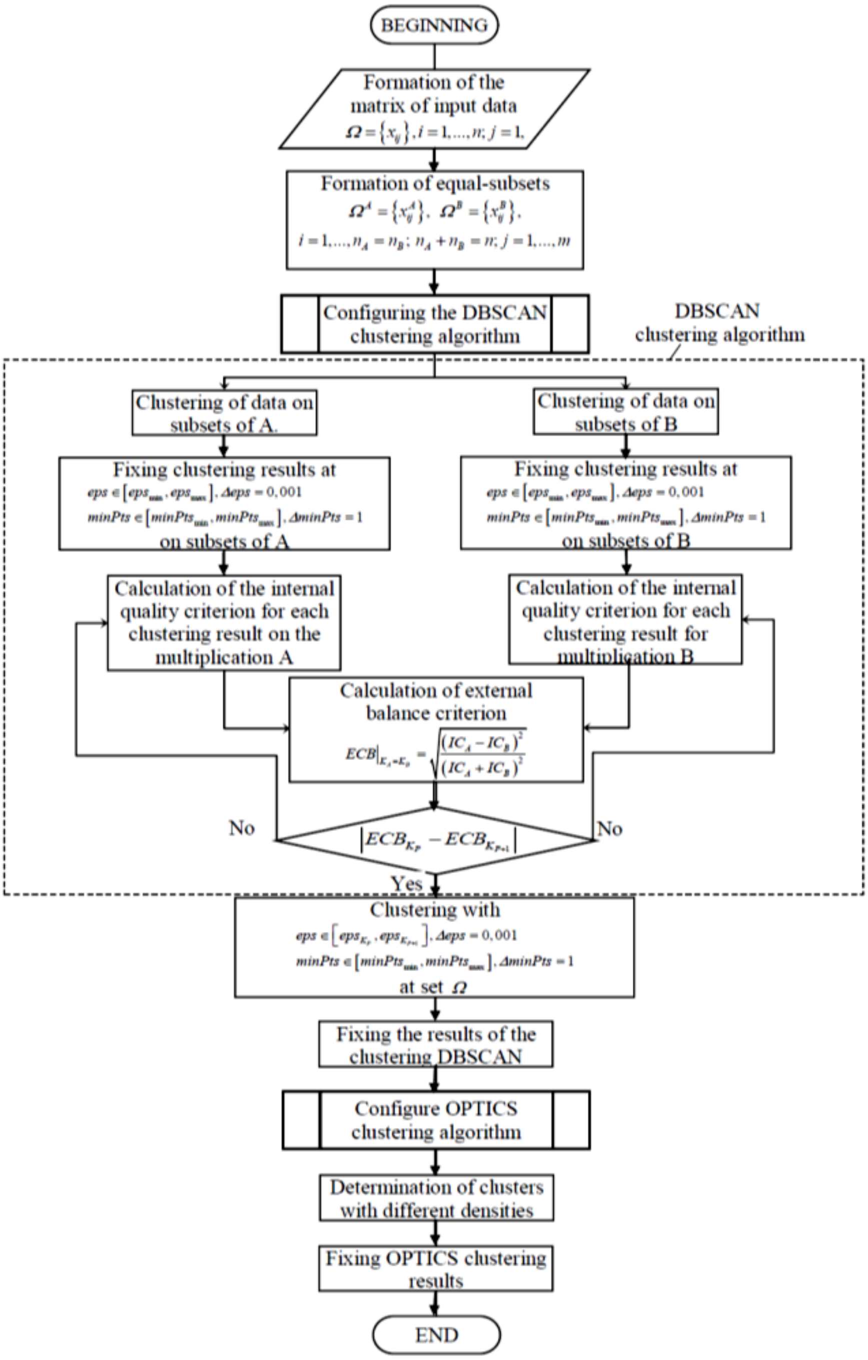
BEGINNING
Formation of the matrix of input data
$\Omega = \{x_{ij}\}, i = 1,...,n; j = 1,$
Formation of equal-subsets
$\Omega^A = \{x_{ij}^A\}, \Omega^B = \{x_{ij}^B\},$
$i = 1,...,n_A = n_B; n_A + n_B = n; j = 1,...,m$
Configuring the DBSCAN clustering algorithm
DBSCAN clustering algorithm
Clustering of data on subsets of A.
Clustering of data on subsets of B
Fixing clustering results at
$eps \in [eps_{min}, eps_{max}], \Delta eps = 0,001$
$minPts \in [minPts_{min}, minPts_{max}], \Delta minPts = 1$
on subsets of A
Fixing clustering results at
$eps \in [eps_{min}, eps_{max}], \Delta eps = 0,001$
$minPts \in [minPts_{min}, minPts_{max}], \Delta minPts = 1$
on subsets of B
Calculation of the internal quality criterion for each clustering result on the multiplication A
Calculation of the internal quality criterion for each clustering result for multiplication B
Calculation of external balance criterion
$ECB|_{K_A=K_B} = \sqrt{\frac{(IC_A - IC_B)^2}{(IC_A + IC_B)^2}}$
No
$|ECB_{K_p} - ECB_{K_{p+1}}|$
No
Yes
Clustering with
$eps \in [eps_{K_p}, eps_{K_{p+1}}], \Delta eps = 0,001$
$minPts \in [minPts_{min}, minPts_{max}], \Delta minPts = 1$
at set Ω
Fixing the results of the clustering DBSCAN
Configure OPTICS clustering algorithm
Determination of clusters with different densities
Fixing OPTICS clustering results
END

Fig. 2. Block diagram of a two-step objective clustering algorithm using DBSCAN and OPTICS algorithms

Step 3. The division Ω into two equally powerful subsets in accordance with the above algorithm. The resulting subsets Ω^A and Ω^B formally can be represented as follows:

$$\begin{gathered}\Omega^A = \left\{x_{ij}^A\right\}, \Omega^B = \left\{x_{ij}^B\right\}, j = 1,...,m \\ i = 1,...,n_A = n_B, n_A + n_B = n\end{gathered} \tag{11}$$

Step 4. Configure the DBSCAN clustering algorithm.

For each equally powerful subset:

Step 5. Data clustering on a subset by the DBSCAN algorithm.

Step 6. Fixing the results of clustering with

$$\begin{gathered}eps \in \left[eps_{\min}, eps_{\max}\right], \Delta eps = 0{,}001 \\ minPts \in \left[minPts_{\min}, minPts_{\max}\right], \Delta minPts = 1\end{gathered} \tag{12}$$

Step 7. Calculation of internal criteria for the quality of clustering for each clustering result.

Step 8. Calculation of the external balance criterion in accordance with the formula (3)

Step 9. If the modulus of the difference in the values of the external balance criteria with the same number of clusters on each subset does not reach the minimum value (4), then Steps 7–8 are repeated.

Otherwise:

Step 10. Fixed clustering algorithm DBSCAN on the set Ω with (5):

Step 11. Setting up the OPTICS clustering algorithm.

Step 12. Identify data clusters with different densities.

Step 13. Fixation of clustering results by the OPTICS algorithm.

Step 14. End

5 Experiment, Results and Discussion

For the first experimentation, the bulletins of the interim test are given on the basis of the comprehensive schools of the federal university [28], which can be folded together. The results of clustering are presented in table 3.

In the second experiment, the algorithms were evaluated using the indices analysis of the results obtained on data that contain clusters of different forms shows that the use of the DBSCAN algorithm of objective clustering inductive technology allows us to adequately group the objects under study. At the same time, the points, the distribution density of which in the feature space is smaller than the distribution density of the objects that make up the clusters, are grouped into a separate cluster.

These points are identified as noise. In accordance with the principles of inductive modeling of complex systems, at the last step, the best solutions are formed that answer (4) for optimal combinations of the algorithm parameters.

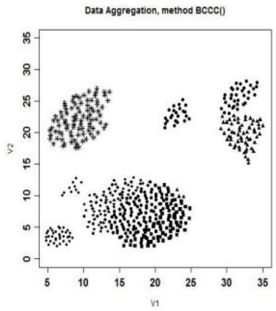

Fig. 3. Data Aggregation: the number of classes is 7; the number of dimension is 2; the number of copies is 788.

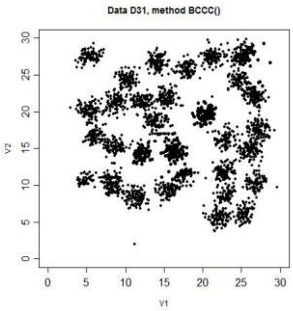

Fig. 4. Data D31: the number of classes is 31; the number of dimension is 2; the number of specimens is 3100.

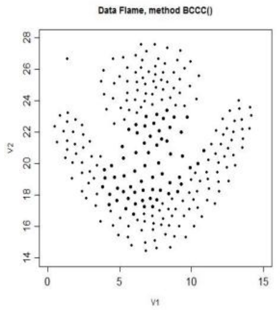

Fig. 5. Data Flame: the number of classes is 2; the number of dimension is 2; the number of copies is 240.

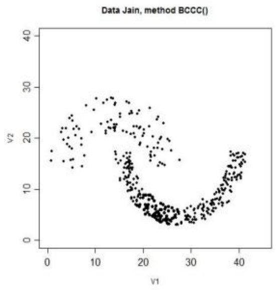

Fig. 6. Data Jain: the number of classes is 2; the number of dimension is 2; the number of copies is 373.

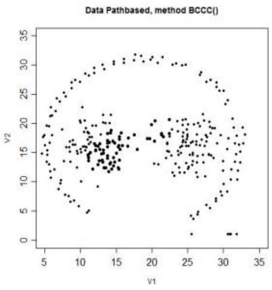

Fig. 7. Data Pathbased: the number of classes is 3; the number of dimension is 2; the number of specimens is 300.

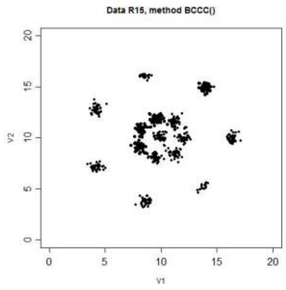

Fig. 8. Data R15: the number of classes is 15; the number of dimension is 2; the number of copies is 600.

Table 3. Results of comparative experiments (percentage of correctly recognized data) of clustering

Data	Inductive DBSCAN+OPTICS	Inductive DBSCAN	DBSCAN	OPTICS
Aggregation	98,5	93,0	87,5	90,1
D31	98,1	93,3	83,4	89,2
Flame	95,7	89,8	79,6	85,4
Jain	98,8	96,2	77,0	92,4
Pathbased	98,2	96,5	81,5	83,2
R15	99,2	97,7	89,2	94,1

The choice of the final solution using the OPTICS algorithm to determine clusters on data with different densities is determined by the goals of the problem being solved. As the results showed (Table 1), the best solutions for choosing the parameters of the DBSCAN algorithm from the point of view of internal criteria are the following: “Aggregation” data: EPS = 0.168, minpts = 4; Compound data: EPS = 0.175, minpts = 4; Iris data: EPS = 0.71, minpts = 3

Thus, we can conclude that the proposed hybrid objective clustering model based on the density algorithm DBSCAN followed by the use of the OPTICS algorithm allows detecting meaningful clusters in data with different densities.

Table 4. Results of clustering quality assessment using Calinski – Harabasz and Dunn’s indices

Data Sets	Inductive DBSCAN+OPTICS		No. of Clusters, $eps, minPts$	No. of Clusters, eps_cl, Xi OPTICS	Inductive DBSCAN		No. of Clusters, $eps, minPts$ DBSCAN
	Index Calinski–Harabasz	Dunn’s index			Index Calinski–Harabasz	Dunn’s index	
Aggregation	534.0648	0.0851	6 $eps = 0.168$ $minPts = 4$	7 $eps\_cl = 0.168$ $Xi = 0.03$	531.6497	0.0866	6 $eps = 0.11$ $minPts = 3$
Compound	535.3819	0.1287	5 $eps = 0.175$ $minPts = 4$	7 $eps\_cl = 0.2$ $Xi = 0.05$	515.001	0.1297	5 $eps = 0.166$ $minPts = 3$
Iris	487.5214	0.1375	5 $eps = 0.71$ $minPts = 3$	7 $eps\_cl = 0.6$ $Xi = 0.03$	468.3452	0.1452	3 $eps = 0.66$ $minPts = 3$

6 Conclusion

The article demonstrates the results of the accomplishment of the objective clustering inductive technology based on the DBSCAN clustering algorithm with the subsequent use of the OPTICS algorithm. The fulfillment of this technology involves the simultaneous clustering of data on two equal power sets, which include the same number of pairs of similar objects.

The external, balance and internal criteria for the quality of clustering were used to determine the studied data objective clustering. The Calinski-Harabasz and Dunn’s criteria were used as an internal quality criterion for clustering.

The external criterion was calculated as the normalized difference of internal quality criteria. At the same time internal quality criteria was calculated on two equal power subsets. The balance criterion was used as an external criterion. The determination of the EPS is the neighborhood and *MinPts* within the values of the EPS is the neighborhood was performed as maximal value of the clustering quality criterion of the complex balance during the operation of the algorithm.

Aggregation D31, Flame, was used as experimental data. Jain, Pathbased, R15, Compound data connections of the Computing School of the East-Finnish University, and well-known also Iris data

The results of simulation showed high efficiency of the proposed technology. In the case of Aggregation and Connections data, the studied objects were adequately divided into clusters. The noise component of the distribution density of objects was selected when the algorithm was running.

Reference

1. Ester, M., Kriegel, H.-P., Xu, X.: Knowledge Discovery in Large Spatial Databases: Focusing Techniques for efficient Class Identification. In: Proceedings of the 4th Int. Symp. on large Spatial Databases, Portland, ME, Vol. 951, 67-82. (1995).
2. Ankerst, M., Breunig, M., Kriegel, H.-P., Sander, J.: OPTICS: Ordering Points To Identify the Clustering Structure. In: Proceedings of Int. Conf. on Management of Data (SIGMOD-99), 49-60. (1999).
3. Ivakhnenko, A.G.: Objective Clustering Based on the Model Self-Organization Theory, Avtomatika, Vol. 5, 6–15. (1987) (in Russian)
4. Jain, A. K., Dubes, R. C. Algorithms for Clustering Data. Prentice-Hall advanced reference series. Prentice-Hall, Inc., Upper Saddle River, NJ. (1988).
5. Nagpal, A., Jatain, A., Gaur, D.: Review based on data clustering algorithms. In: Proceeding of the IEEE Conference, Information & Communication Technologies (ICT), 298-303. (2013)
6. Ester, M., Kriegel, H., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proc. Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press, Vol. 96/34, 226-231. (1996).
7. Sarmah, S., Bhattacharyya, D.: A grid-density based technique for finding clusters in satellite image, Vol. 33/5, 589-604. (2012).
8. Ankerst, M., Breunig, M., Kriegel, H., Sander, J.: OPTICS: ordering points to identify the clustering structure. In Proc. ACM SIGMOD international conference on Management of data, Vol. 28/2, 49-60. (1999)
9. Madala, H.R., Ivakhnenko, A.G.: Inductive Learning Algorithms for Complex Systems Modeling. In: CRC Press Inc., Boca Raton, 365 p. (1994)
10. Stepashko, V., Bulgakova, O., Zosimov, V.: Construction and research of the generalized iterative GMDH algorithm with active neurons. In: Advances in Intelligent Systems and Computing II, 492–510. (2018). DOI: 10.1007/978-3-319-70581-1_35
11. Bulgakova, O., Stepashko, V., Zosimov, V.: Numerical study of the generalized iterative algorithm GIA GMDH with active neurons. In: Proceedings of the 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies, Vol.1, art. No. 8098836, 496–500. (2017). DOI: 10.1109/STC-CSIT.2017.8098836
12. Osypenko, V. V., Reshetjuk, V.M.: The methodology of inductive system analysis as a tool of engineering researches analytical planning. In: Ann. Warsaw Univ. Life Sci, 2011, SGGW, Vol. 58, 67–71.(2011). [Electronic resource]. – Access mode: http://annals-wuls.sggw.pl/?q=node/234

13. Lurie, I.A., Osipenko, V.V., Litvinenko, V.I., Taif, M.A., Kornilovska, N.V.: Hybridization of the algorithm of inductive cluster analysis using estimation of data distribution. In: Lviv Polytechnic: Information systems and networks, Vol. 832, 178-190. (2015) (in Ukrainian).
14. Babichev, S., Taif, M., Lytvynenko, V., Korobchinskyi, M. :Objective clustering inductive technology of gene expression sequences features, Communications in Computer and Information Science: In the book "Beyond Databases, Architectures and Structures", 359–372. (2017).
15. Babichev, S., Lytvynenko, V., Osypenko, V.: Implementation of the objective clustering inductive technology based on the DBSCAN clustering algorithm. In: Proceeding of the XIIth IEEE international scientific and technical conference, 479-484. (2017).
16. Babichev, S., Vyshemyrska, S., Lytvynenko, V.: Implementation of DBSCAN Clustering Algorithm within the Framework of the Objective Clustering Inductive Technology based on R and KNIME. In: Radio Electronics, Computer Science, Control-2019, No.1, 77-88. (2019)
17. Ester, M., Kriegel, H.-P., Sander, J., Xu, X.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining. Portland, OR, 226-231.(1996).
18. Bäcklund, H., Hedblom, A., Neijman, N.: A Density-Based Spatial Clustering of Application with Noise. Linköpings Universitet. (2011).
19. Son, T. M.: Density-based algorithms for active and anytime clustering. In: Ludwig Maximilians University Munich (2014).
20. Ivakhnenko, A.G.: Heuristic Self-Organization. In: Problems of Engineering Cybernetics. Automatica, No. 6, 207-219. (1970).
21. Gabor, D.: Planning Perspectives. Automation, No.2, 16-22. (1972) (in Russian)
22. Stepashko, V.C.: Elements of the theory of inductive modeling. The state and prospects of the development of computer science in Ukraine: monograph. Kyiv: Scientific Opinion, 471-486. (2010). (in Ukrainian)
23. Zholnarsky, A.A.: Agglomerative Cluster Analysis Procedures for Multidimensional Objects: A Test for Convergence. In: Pattern Recognition and Image Analysis, Vol.25, No.4, 389-390. (1992).
24. Ivakhnenko, A.G., Ivakhnenko, G.A.: The Review of Problems Solvable by Algorithms of the Group Method of Data Handling (GMDH). In: Pattern Recognition and Image Analysis, Vol.5, No.4, 527-535. (1995)
25. Bezdek, J.C., Dunn, J.C.: Optimal fuzzy partitions: A heuristic for estimating the parameters in a mixture of normal distributions. In: Proceeding of the IEEE Transactions on Computers, 835–838. (1975)
26. Calinski, R.B., Harabasz, J.: A dendrite method for cluster analysis. In: Comm. in Statistics, Vol. 3:1, 27p. (1974).
27. Babichev, S., Taif, M., Lytvynenko, V.: Inductive model of data clustering based on the agglomerative hierarchical algorithm. In: Proceeding of the 2016 IEEE First International Conference on Data Stream Mining and Processing (DSMP), 19– 22. (2016). [Electronic resource]. – Access mode: http://ieeexplore.ieee.org/document/7583499/
28. https://cs.joensuu.fi/sipu/datasets/