ISSN 8756-6990, Optoelectronics, Instrumentation and Data Processing, 2011, Vol. 47, No. 3, pp. 245–252. © Allerton Press, Inc., 2011.
Original Russian Text © I.A. Pestunov, V.B. Berikov, E.A. Kulikova, S.A. Rylov, 2011, published in Avtometriya, 2011, Vol. 47, No. 3, pp. 49–58.

ANALYSIS AND SYNTHESIS OF SIGNALS AND IMAGES

Ensemble of Clustering Algorithms for Large Datasets

I. A. Pestunov[a], V. B. Berikov[b], E. A. Kulikova[a], and S. A. Rylov[a]

[a] *Institute of Computational Technologies, Siberian Branch, Russian Academy of Sciences, pr. Akademika Lavrent'eva 6, Novosibirsk, 630090 Russia*
E-mail: pestunov@ict.nsc.ru
[b] *Sobolev Institute of Mathematics, Siberian Branch, Russian Academy of Sciences, pr. Akademika Koptyuga 4, Novosibirsk 630090, Russia*

Received April 11, 2011

Abstract—The ensemble clustering algorithm ECCA (Ensemble of Combined Clustering Algorithms) for processing large datasets is proposed and theoretically substantiated. Results of an experimental study of the algorithm on simulated and real data proving its effectiveness are presented

Keywords: ensemble clustering algorithm, grid-based approach, large datasets.

DOI: 10.3103/S8756699011030071

INTRODUCTION

In recent years, the efforts of many researchers were focused on the creation of efficient clustering algorithms to analyze large datasets (genetic data, multispectral images, Internet data, etc.) [1, 2]. The demand for such algorithms is continuously increasing due to the rapid progress in the creation of means and technologies of automated acquisition and storage of data, and due to the fast development of Internet technologies.

One of the most effective approaches to clustering large datasets is the so-called grid-based approach [3], which involves transition from clustering of individual objects to clustering of the elements of the grid structure (cells) formed in an attribute space. This approach assumes that all objects that fell in the same cell belong to the same cluster. Therefore, the formation of the grid structure is an important step in the algorithm.

According to the methods of constructing the grid structure, the clustering algorithms can conventionally be divided into two groups [4]: algorithms with an adaptive grid and with a fixed grid.

The algorithms with an adaptive grid analyze the data distribution in order to make the most accurate description of the boundaries of the clusters formed by the original objects [5]. In an adaptive grid, the grid (boundary) effect is reduced, but its construction, as a rule, involves significant computing costs.

Algorithms with a fixed grid have a high computing efficiency, but the clustering quality in most cases is low because of the grid effect, and the obtained results are unstable because they depend on the scale of the grid. In practice, this instability makes it difficult to configure the parameters of the algorithm.

To solve this problem, grid-based methods which use not one but several grids with a fixed pitch have been actively developed in recent years, [6–8]. The main difficulties of this approach are in the development of a method for combining the results obtained on different grids because the formed clusters are not always clearly comparable with each other. In [6] an algorithm is presented which performs clustering on a sequence of grids until a repetitive (stable) result is obtained. In the algorithms of [7, 8], two clustering operations are performed on grids of different sizes. The final result is formed by combining the overlapping clusters constructed on each of these grids.

In the present paper, to improve the quality and stability of solutions, we propose the clustering algorithm ECCA (Ensemble of Combined Clustering Algorithms) which uses an ensemble of algorithms with fixed uniform grids and in which the final collective solution is based on pairwise classification of the elements of the grid structure.

1. FORMULATION OF THE PROBLEM

Let the set of objects X being classified consist of vectors lying in the attribute space R^d: $X = \{x_i = (x_i^1, \ldots, x_i^d) \in R^d,\ i = \overline{1, N}\}$. The vectors x_i lie in a rectangular hyperparallelepiped $\Omega = [l^1, r^1] \times \ldots \times [l^d, r^d]$, where $l^j = \min\limits_{x_i \in X} x_i^j$ and $r^j = \max\limits_{x_i \in X} x_i^j$. Under the grid structure we mean a partition of the attribute space by hyperplanes $x^j = (r^j - l^j)i/m + l^j$, $i = 0, \ldots, m$ (m is the number of partitioned areas in each dimension). The minimum element of this structure is a cell (a closed rectangular hyperparallelepiped bounded by hyperplanes). Let us introduce a common numbering of the cells (sequentially from one layer of cells to another).

The cells B_i and B_j ($i \neq j$) are adjacent if their intersection is not empty. The set of cells adjacent to B will be denoted as A_B. By the density D_B of the cell B we mean the ratio $D_B = N_B/V_B$, where N_B is the number of elements of the set X that fell in the cell B; V_B is the volume of the cell B. We assume that the cell B is non-empty if $D_B \geq \tau$, where τ is the magnitude of the specified threshold. All points of the set X that fell in the cells with a density less than τ will be classified as noise. Let us denote the set of all non-empty cells as $\aleph$. The non-empty cell B_i is directly connected to the non-empty cell B_j ($B_i \to B_j$) if B_j is the cell with the maximum number that satisfies the conditions $B_j = \arg\max\limits_{B_k \in A_{B_i}} D_{B_k}$ and $D_{B_j} \geq D_{B_i}$.

The non-empty cells B_i and B_j are directly connected ($B_i \rightleftarrows B_j$) if $B_i \to B_j$ or $B_j \to B_i$. The non-empty cells B_i and B_j are connected to each other ($B_i \sim B_j$) if there exist $k_1, \ldots, k_l$ such that $k_1 = i$, $k_l = j$ and for all $p = 1, \ldots, l-1$, we have $B_{k_p} \rightleftarrows B_{k_{p+1}}$.

The introduction of relationship of connectedness causes a natural partition of the nonempty cells into the connectedness components $\{G_i,\ i = 1, \ldots, S\}$. By the connectedness component we mean the maximum set of pairwise connected cells. The cell $Y(G)$ satisfying the condition $Y(G) = \arg\max\limits_{B \in G} D_B$ will be called a representative of the connectedness component G [if several cells satisfy this condition, then $Y(G)$ is selected from them randomly]. The connectedness components G' and G'' are adjacent if there exist adjacent cells B' and B'' such that $B' \in G'$ and $B'' \in G''$. The adjacent connectedness components of G_i and G_j are connected ($G_i \sim G_j$) if there exists a set of cells (path) $P_{ij} = \{Y_i = B_{k_1}, \ldots, B_{k_t}, \ldots, B_{k_l} = Y_j\}$ such that:

1) for all $t = 1, \ldots, l-1$, the cell $B_{k_t} \in G_i \cup G_j$ and B_{k_t} and $B_{k_{t+1}}$ are adjacent cells;

2) $\min\limits_{B_{k_t} \in P_{ij}} D_{B_{k_t}} / \min(D_{B_{Y_i}}, D_{B_{Y_j}}) > T$, $T > 0$ is the grouping threshold.

Definition. The maximum set of pairwise connected connectedness components will be called a cluster C: 1) for all connectedness components $G_i, G_j \in C$, the relation $G_i \sim G_j$ holds; 2) for any $G_i \in C$, $G_j \notin C$, the relation $G_i \not\sim G_j$ holds.

In view of the foregoing, the clustering problem is to partition the set $\aleph$ into an ensemble of clusters $\{C_i,\ i = 1, \ldots, M\}$ such that $\aleph = \bigcup\limits_{i=1}^{M} C_i$ and $C_i \cap C_j = \varnothing$ for $i \neq j$; the number of clusters M is not known beforehand.

Next, we describe an efficient method for solving this problem based on an ensemble approach.

2. DESCRIPTION OF THE METHOD

The propsoed method is based on the CCA(m, T, τ) grid-based algorithm [9], where m is the number of partitions, T is the threshold of grouping of the connectedness components, and τ is the noise threshold. This algorithm can be divided into three main steps.

1. Formation of the cell structure. In this step, for each point $x_i \in X$, the cell containing it is determined, the densities D_B of all cells are calculated, and non-empty cells are identified.

2. Isolation of the connectedness components of $G_1, \ldots, G_S$ and search for their representatives $Y(G_1), \ldots, Y(G_S)$.

3. Formation of clusters $C_1, \ldots, C_M$ in accordance with the above definition based on the isolated connectedness components.

The CCA(m, T, τ) algorithm is computationally efficient in the attribute space of small dimension ($\leqslant 6$) [9], its complexity is $O(dN + dm^d)$, where N is the number of classified objects, d is the dimension of the attribute space.

However, the CCA belongs to the class of fixed-grid algorithm; thereore, the results of its work greatly depend on the parameter m which determines the scale of the elements of the grid structure. In practice, this instability of results considerably complicates the configuration of parameters of the algorithm.

It is known [10–12] that the stability of solutions in clustering problems can be increased by the formation of an ensemble of algorithms and construction of a collective solution on its basis. This is done using the results obtained by different algorithms or the same algorithm with different values of parameters. In addition, various subsystems of variables can be applied to the formation of an ensemble. The ensemble approach is one of the most promising trends in cluster analysis [1].

In this paper, it is suggested that an ensemble is formed using the results of implementation of the CCA(m, T, τ) algorithm with different values of the parameter m and the final collective solution is obtained by applying the method based on finding a consistent similarity matrix (or differences) of objects [13]. This method can be described as follows.

Suppose that using a certain clustering algorithm [$\mu = \mu(\Theta)$] which depends on a random parameter vector $\Theta \in \mathbf{\Theta}$ ($\mathbf{\Theta}$ is an admissible set of parameters), we obtained a set of partial solutions $\mathbb{Q} = \{Q^{(1)}, \ldots, Q^{(l)}, \ldots, Q^{(L)}\}$, where $Q^{(l)}$ is the lth version of the clustering which contains $M^{(l)}$ clusters.

We use $H(\Theta_l)$ to denote an $N \times N$ binary matrix $H(\Theta_l) = \{H_{i,j}(\Theta_l)\}$, which for the lth group is introduced as

$$H_{i,j}(\Theta_l) = \begin{cases} 0 & \text{if objects are grouped into the same cluster;} \\ 1 & \text{otherwise.} \end{cases}$$

After the construction of L partial solutions, it is possible to form a consistent matrix of differences

$$\mathbf{H} = \{\mathbf{H}_{i,j}\}, \quad \mathbf{H}_{i,j} = \frac{1}{L}\sum_{l=1}^{L} H_{i,j}(\Theta_l),$$

where $i, j = 1, \ldots, N$. The quantity $\mathbf{H}_{i,j}$ equals the frequency of classification of x_i and x_j into different groups in the set of groups $\mathbb{Q}$. A value of this quantity close to zero implies that these objects have a great chance of falling into the same group. A value of this quantity close to unity indicates that the chance of falling in the same group is negligible for these objects.

In our case, $\mu = \text{CCA}(m, T, \tau)$, where the number of partitions $m \in \{m_{\min}, m_{\min}+1, \ldots\ldots, m_{\min}+L\}$, and the objects of classification will be representatives of the connectedness components $Y(G_1), \ldots, Y(G_S)$.

After calculating the consistent matrix of differences, to obtain a collective solution, we apply the standard agglomerative method of dendrogram construction which uses pairwise distances between objects as input data [14]. The distances between the groups will be determined in accordance with the principle of mean connection, i.e., as the arithmetic mean of the pairwise distances between the objects included in the groups. The grouping process continues until the distance between the closest groups exceeds the specified threshold value T_d belonging to the interval [0, 1]. This method highlights the hierarchical structure of the clusters, which simplifies the interpretation of the results.

3. THEORETICAL BASIS OF THE METHOD

To investigate the properties of the proposed method of forming a collective solution, we consider a probabilistic model.

Suppose that there is a hidden (directly unobservable) variable U which specifies the classification of each object into some of M classes (clusters). We consider the following probabilistic model of data generation. Suppose that each class has a specific law of conditional distribution $p(x \,|\, U = = i) = p_i(x)$, where $x \in R^d$ and $i = 1, \ldots, M$. For each object we determine the class into which it falls in accordance with the a priori probabilities $P_i = P(U = i)$ $(i = 1, \ldots, M)$, where $\sum\limits_{i=1}^{M} P_i = 1$,. Then, the observed value of x is calculated in accordance with the distribution $p_i(x)$. This procedure is performed independently for each object, and the result is a random sample of objects.

Suppose that set of objects is partitioned into M subsets using a certain cluster analysis algorithm μ. Since the numbering of the clusters is not important, it is more convenient to consider the equivalence relation, i. e., to indicate whether the algorithm μ places each pair of objects in the same class or in different classes. We consider a random pair a, b of different objects and define the quantity

$$h_{\mu,a,b} = \begin{cases} 0 & \text{if objects are placed in the same class;} \\ 1 & \text{otherwise.} \end{cases}$$

Let $P_U = P(U(a) \neq U(b))$ be the probability that the objects belong to different classes. The probability of the error that can be made by the algorithm μ in the classification of a and b will be denoted by $P_{\text{err},\mu}$, where

$$P_{\text{err},\mu} = \begin{cases} P_U & \text{if } h_{\mu,a,b} = 0, \\ 1 - P_U, & \text{if } h_{\mu,a,b} = 1. \end{cases}$$

It is easy to see that

$$P_{\text{err},\mu} = P_U + (1 - 2P_U)h_{\mu,a,b}. \tag{1}$$

Suppose the algorithm μ depends on the random parameter vector $\Theta \in \mathbf{\Theta}$: $\mu = \mu(\Theta)$. To emphasize the dependence of the algorithm results on the parameter Θ, from now we will denote $h_{\mu(\Theta),a,b} = h(\Theta)$ and $P_{\text{err},\mu(\Theta)} = P_{\text{err}}(\Theta)$.

Suppose that a set of solutions $\theta_1, \ldots, \theta_L$ was obtained as a result of L-fold application of the algorithm μ with random and independently selected parameters $h(\theta_1), \ldots, h(\theta_L)$. For the sake of definiteness, we assume that L is odd. The function

$$H(h(\theta_1), \ldots, h(\theta_L)) = \begin{cases} 0 & \text{if } \dfrac{1}{L}\displaystyle\sum_{l=1}^{L} h(\theta_l) < \dfrac{1}{2}; \\ 1 & \text{otherwise} \end{cases}$$

will be called the collective (ensemble) solution. It is necessary to investigate the behavior of the collective solution as a function of the size of the ensemble L. Note that each individual algorithm can also be regarded as a degenerate case of the ensemble with $L = 1$.

Proposition 1. The initial moment of the kth order for the error probability of the algorithm $\mu(\Theta)$ equals

$$\nu_k = (1 - P_h)P_U^k + P_h(1 - P_U)^k,$$

where $P_h = P(h(\Theta) = 1)$.

Proof. The validity of the expression follows from the fact that

$$\nu_k = \mathbf{E}_\Theta P_{\text{err}}^k(\Theta) = \mathbf{E}_\Theta(P_U + (1 - 2P_U)h(\Theta))^k = \mathbf{E}_\Theta \sum_{m=0}^{k} C_k^m P_U^m (1 - 2P_U)^{k-m} h^{k-m}(\Theta)$$

$$= \sum_{m=0}^{k} C_k^m P_U^m (1 - 2P_U)^{k-m} \mathbf{E}_\Theta h^{k-m}(\Theta).$$

Since $\mathbf{E}_\Theta h^q(\Theta) = \mathbf{E}_\Theta h(\Theta) = P_h$ for $q > 0$, we obtain

$$\nu_k = P_U^k + \sum_{m=1}^{k} C_k^m P_U^m (1 - 2P_U)^{k-m} P_h = P_U^k - P_h P_U^k + P_h \sum_{m=0}^{k} C_k^m P_U^m (1 - 2P_U)^{k-m}$$

$$= P_U^k - P_h P_U^k + P_h(P_U + 1 - 2P_U)^k = P_U^k - P_h P_U^k + P_h(1 - P_U)^k = (1 - P_h)P_U^k + P_h(1 - P_U)^k,$$

which was to be proved.

Corollary 1. The mathematical expectation and variance of the error probability for the algorithm $\mu(\Theta)$ are equal, respectively, to

$$\mathbf{E}_\Theta P_{\text{err}}(\Theta) = P_U + (1 - 2P_U)P_h, \quad \mathbf{Var}_\Theta P_{\text{err}}(\Theta) = (1 - 2P_U)^2 P_h(1 - P_h).$$

Proof. Validity of the expression for the mathematical expectation follows from the proved proposition for the moment ν_1 and directly from (1). Let us consider the expression for the variance. According to the definition

$$\mathbf{Var}_\Theta P_{\text{err}}(\Theta) = \nu_2 - \nu_1^2.$$

Hence,

$$\mathbf{Var}_\Theta P_{\text{err}}(\Theta) = (1 - P_h)P_U^2 + P_h(1 - P_U)^2 - (P_U + (1 - 2P_U)P_h)^2.$$

After transformations, we obtain

$$\mathbf{Var}_{\Theta} P_{\text{err}}(\Theta) = (1 - 2P_U)^2 P_h(1 - P_h),$$

which was to be proved.

We denote by $P_{\text{err}}(\Theta_1, \ldots, \Theta_L)$ a random function whose equals the probability of the error that can be made by the ensemble algorithm in the classification of a and b. Here $\Theta_1, \ldots, \Theta_L$ are independent statistical copies of the random vector Θ. Consider the behavior of the error probability for the collective solution.

Proposition 2. The initial moment of the kth order for the error probability of the collective solution is

$$\mathbf{E}_{\Theta_1, \ldots, \Theta_L} P^k_{\text{err}}(\Theta_1, \ldots, \Theta_L) = (1 - P_{H,L})P^k_U + P_{H,L}(1 - P_U)^k,$$

where

$$P_{H,L} = P\Big(\frac{1}{L}\sum_{l=1}^{L} h(\Theta_l) \geq \frac{1}{2}\Big) = \sum_{l = \lfloor L/2 \rfloor + 1}^{L} C^l_l P^l_h (1 - P_h)^{L-1}$$

($\lfloor \cdot \rfloor$ denotes the integer part).

The proof of this proposition is similar to the proof of Proposition 1 [the error probability of the collective solution is determined by a formula similar to formula (1)]. Moreover, it is clear that the distribution of the number of votes given for the solution $h = 1$ is binomial: $\text{Bin}(L, P_h)$.

As in Proposition 1, it is possible to show that the mathematical expectation and variance of the error probability for the collective solutions are equal, respectively, to

$$\mathbf{E}_{\Theta_1, \ldots, \Theta_L} P_{\text{err}}(\Theta_1, \ldots, \Theta_L) = P_U + (1 - 2P_U)P_{H,L},$$

$$\mathbf{Var}_{\Theta_1, \ldots, \Theta_L} P_{\text{err}}(\Theta_1, \ldots, \Theta_L) = (1 - 2P_U)^2 P_{H,L}(1 - P_{H,L}).$$

Let us use the following a priori information on the cluster analysis algorithm. We assume that the expected probability of misclassification $\mathbf{E}_{\Theta} P_{\text{err}}(\Theta) << 1/2$, i. e., it is assumed that the algorithm μ performs better in the classification than the algorithm of random equiprobable choice. Corollary 1 implies that one of two variants holds: (a) $P_h > 1/2$ and $P_U > 1/2$; (b) $P_h < 1/2$ and $P_U < 1/2$. For definiteness, we consider the first case.

Proposition 3. If $\mathbf{E}_{\Theta} P_{\text{err}}(\Theta) < 1/2$, and thus $P_h > 1/2$ and $P_U > 1/2$, then with increasing power (number of elements) of the ensemble, the expected probability of misclassification decreases tending in the limit to $1 - P_U$, and the variance of the error probability tends to zero.

Proof. The de Moivre–Laplace integral theorem implies that with increasing L,

$$P_{H,L} = 1 - P\Big(\frac{1}{L}\sum_{l=1}^{L} h(\Theta_l) < \frac{1}{2}\Big)$$

converges to

$$1 - \Phi\Big(\frac{1/2 - P_h}{\sqrt{P_h(1 - P_h)/L}}\Big),$$

where $\Phi(\cdot)$ is a distribution function of the standard normal law. Hence, as $L \to \infty$, the value of $P_{H,L}$ increases monotonically tending to unity. The relation

$$E_{\Theta_1, \ldots, \Theta_L} P_{\text{err}}(\Theta_1, \ldots, \Theta_L) = P_U + (1 - 2P_U)P_{H,L},$$

where $(1 - 2P_U) < 0$, and Proposition 2 implies the validity of Proposition 3.

It is obvious that in the latter case, the expected error probability also decreases with increasing power of the ensemble, tending to the quantity P_U, while the error variance tends to zero.

The proved proposition suggests that when the abovementioned natural conditions are satisfied, the application of the ensemble makes it possible to improve the quality of clustering.

Results of Operation of the ECCA Algorithm on Data for Irises

Parameters	Classes		
	$i = 1$	$i = 2$	$i = 3$
$\lvert C_i^O \rvert$	50	50	50
$\lvert C_i^S \rvert$	50	52	48
$\lvert C_i^O \cap C_i^S \rvert$	50	48	46
Accuracy, %	100	96	92
Measure of coverage, %	100	92.31	95.83

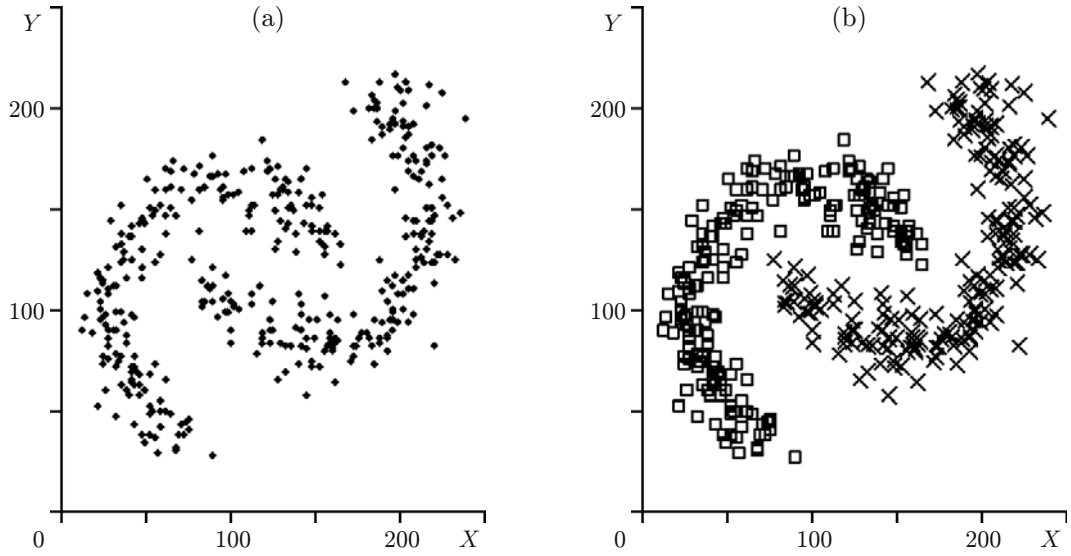

Fig. 1.

4. RESULTS OF EXPERIMENTAL STUDIES

In accordance with the method proposed in Sec. 2, the ECCA($m_{\min}, L, T, \tau, T_d$) ensemble algorithm was developed and implemented in Java. The algorithm requires the specification of values for five parameters: $m_{\min}, L, T, \tau, T_d$. Numerous experimental studies performed on simulated and real data showed that, with the use of ten elements of the ensemble, the obtained results are stable to the choice of the grid parameter $m_{\min}$. The parameter T has a weak effect on the clustering result. In the image processing, this parameter was chosen to be equal to 0.8 and the noise threshold $\tau \in \{0; 1\}$. The ECCA algorithm allows obtaining the hierarchical data structure. The studies show that the parameter T_d specifying the depth of the hierarchy can be chosen from the set $\{0, 0.1, \ldots, 0.9\}$. Below we give the results of experiments performed on simulated and real data and confirming the efficiency of the proposed algorithm. The processing was carried out on a PC with a 3 GHz clock frequency.

Experiment No. 1. The well-known iris setosa data set [15] was used. The set consisted of 150 points of a four-dimensional attribute space grouped into three classes of 50 points. We denote by $|C_i^O|$ the actual number of points of the ith class, and by C_i^O the number of points of the class $|C_i^S|$ contained in the corresponding cluster selected by the ECCA algorithm. Similarly [4], the accuracy of clustering and the measure of coverage by clusters C_i^S of the classes C_i^O will be determined by the formulas $|C_i^O \cap C_i^S|/|C_i^S|$ and $|C_i^O \cap C_i^S|/|C_i^O|$ respectively, where $|\cdot|$ is the cardinality of the set. Table 1 shows the values of the calculated criteria after the application of the ECCA algorithm with the parameters $m_{\min} = 25$, $L = 10$, $T = 0.9$, $\tau = 0$, and $T_d = 0.7$. By these criteria, the results of the ECCA algorithm are superior to the results of the GCOD algorithm [4].

Experiment No. 2. The experiment was performed with two-dimensional data consisting of 400 points grouped into two linearly inseparable classes in the shape of banana (Fig. 1a; the original set). The model

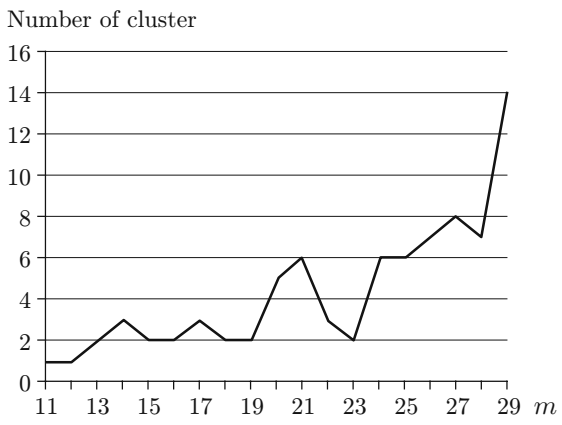

Fig. 2.

Fig. 3.

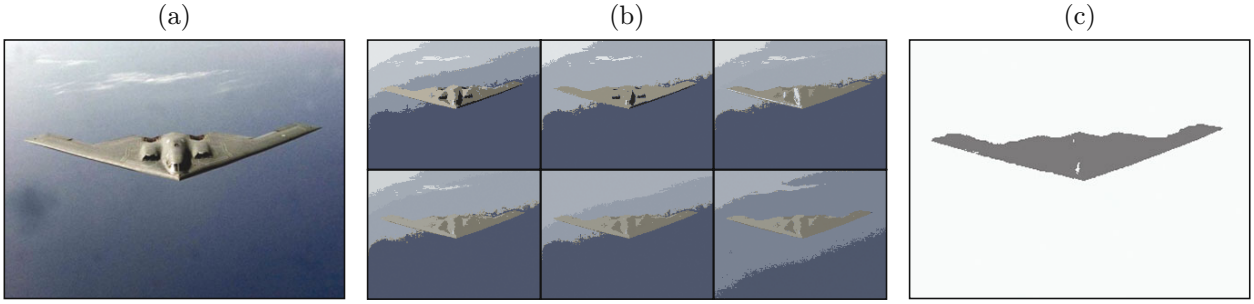

Fig. 4.

was constructed using PRTools (The Matlab Toolbox for Pattern Recognition: http://www.prtools.org) with a parameter of 0.7. Figure 1b shows the results of the ECCA algorithm (15, 10, 0.3, and 0.8). For comparison, the initial data were processed by the DBSCAN algorithm [16]. After a lengthy configuration of its parameters, the result shown in Fig. 1b was achieved. However, the processing time was more than 100 longer than that in with ECCA.

Figure 2 shows a curve of the dependence of the number of clusters obtained by the CCA$(m, 0.8, 0)$ algorithm versus the parameter m which determines the size of the elements of the cell structure. Figure 3 shows a curve of the clustering error versus the values of the parameters m for fixed parameters T and τ for the CCA algorithm (dashed curve) and $m_{\min}$ for fixed parameters T and L for ECCA (solid curve). Here the clustering error is determined by the formula $\sum_{i=1}^{2} |C_i^O \backslash C_i^S| \Big/ \sum_{i=1}^{2} |C_i^O|$. The curves show a substantial dependence of the results of the CCA algorithm on the configurable parameter m and the stability of the obtained solutions for the ECCA ensemble algorithm with variation in the parameter $m_{\min}$. This stability significantly simplifies the configuration of the parameters of the ECCA algorithm.

Experiment No. 3. A 640 × 480 pixel color image (Fig. 4a) (http://commons.wikimedia.org/wiki/File:B-2_Spirit_4.jpg) was processed. Clustering was carried out in the RGB color space. Each cluster corresponded to a homogeneous region in the image. An ensemble of ten elements was used. None of them allows one to identify the object of interest on the original image (Fig. 4b shows six elements of ten). Figure 4c presents the result of applying the ECCA ensemble algorithm with parameters $m_{\min} = 30$, $L = 10$, $T = 0.8$, $\tau = 0$, and $T_d = 0.9$. The processing time was 0.88 s.

CONCLUSIONS

A method for clustering large datasets on the basis of an ensemble of grid-based algorithms was proposed. Its theoretical substantiation is given.

The principal characteristics of the algorithm that distinguish it from other algorithms of cluster analysis are: 1) universality (the algorithm allows one to identify clusters differing in size, shape, and density in the presence of noise); 2) high performance in the clustering of a large number of objects ($\sim 10^6$) (provided that number of attributes is small (≤ 6), this condition is satisfied, in particular, in image analysis problems); 3) ease of parameter configuration.

The results of the experiments performed on simulated and real data confirm the high quality of the obtained solutions and their stability to changes in the configurable parameters. The possibility of obtaining a hierarchical system of the nested clusters greatly facilitates the process of interpretation of results. The high performance of the ECCA algorithm allows interactive processing of large datasets. The ECCA algorithm allows paralleling which increases its performance on multiprocessor systems.

This work was supported by the Russian Foundation for Basic Research (Grants No. 11-07-00346-a, No. 11-07-00202-a).

REFERENCES

1. A. K. Jain, "Data Clustering: 50 Years Beyond K-Means," Pattern Recogn. Lett. **31** (8), 651–666 (2010).
2. D. P. Mercer, *Clustering Large Datasets* (Linacre College, 2003); http://www.stats.ox.ac.uk/∼mercer/documents/Transfer.pdf (date accessed: 03.21.2011).
3. M. R. Ilango and V. Mohan, "A Survey of Grid Based Clustering Algorithms," Int. J. Eng. Sci. Technol. **2**, No. 8, 3441–3446 (2010).
4. B.-Z. Qiu, X.-L. Li, and J.-Y. Shen, "Grid-Based Clustering Algorithm Based on Intersecting Partition and Density Estimation," Lect. Notes Artif. Intel. **4819**, 368–377 (2007).
5. M.-I. Akodjènou-Jeannin, K. Salamatian, and P. Gallinari, "Flexible Grid-Based Clustering," Lect. Notes Artif. Intel. **4702**, 350–357 (2007).
6. W. M. Ma Eden and W. S. Chow Tommy, "A New Shifting Grid Clustering Algorithm," Pattern Recogn. **37**, No. 3, 503–514 (2004).
7. N. P. Lin, C.-I. Chang, H.-E. Chueh, et al., "A Deflected Grid-Based Algorithm for Clustering Analysis," WSEAS Trans. Comput. **7**, No. 4, 125–132 (2008).
8. Y. Shi, Y. Song, and A. Zhang, "A Shrinking-Based Approach for Multi-Dimensional Data Analysis," in *Proc. of the 29th VLDB Conf.* (Berlin, Germany, 2003), pp. 440–451.
9. E. A. Kulikova, I. A. Pestunov, and Y. N. Sinyavskii, "Nonparametric Clustering Algorithm for Large Datasets," in *Proc. of 14 Nat. Conf. "Mathematical Methods for Pattern Recognition"* (MAKS Press, Moscow, 2009), pp. 149–152.
10. A. Strehl and A. Ghosh, "Clustering Ensembles — A Knowledge Reuse Framework for Combining Multiple Partitions," J. Mach. Learn. Res. **3**, 583–617 (2002).
11. A. S. Biryukov, V. V. Ryazanov, and A. S. Shmakov, "Solution of Cluster Analysis Problems Using Groups of Algorithms," Zh. Vychisl. Mat. Mat. Fiz. **48**, No. 1, 176–192 (2008).
12. Y. Hong and S. Kwong. "To Combine Steady-State Genetic Algorithm and Ensemble Learning for Data Clustering," Pattern Recogn. Lett. **29**, No. 9, 1416–1423 (2008).
13. V. B. Berikov, "Construction of the Ensemble of Logical Models in Cluster Analysis," Lect. Notes Artif. Intel. **5755**, 581–590 (2009).
14. R. Duda and P. Hart, *Pattern Recognition and Scene Analysis* (Wiley, New York, 1973).
15. M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics* (London, Charles Cliffin, 1968).
16. M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Database," in *Proc. of the Int. Conf. Knowledge Discovery and Data Mining* (1996), pp. 226–231.