The Hybridization of Inductive Algorithm of Cluster Analysis with the Use of Assessment of Density Distribution Data

[1]Lur'ye I., [2]Osypenko V., [1]Lytvynenko V.

[1]*Kherson National Technical University*
[2]*National University of Life and Environmental Sciences of Ukraine*

iil@rambler.ru, vvo7@ukr.net, immun56@gmail.com

Abstract. *In this paper at the decision of applied problems in the field of bioinformatics data processing has been proposed the inductive clustering algorithm with built-in algorithm of DBSCAN (Density Based Spatial Clustering of Applications with Noise), in which decisions are approximate to global minimum that can be received by successive run of DBSCAN operations.*

Keywords

Inductive modeling, clusterization, external of unbiasedness criterion

1 Introduction

The task of clusterization is a special case of training task “without a teacher” that is reduced to partitioning the available set of data objects into subsets so that the elements of one subset has significantly different of elements properties from all other subsets of objects [1-3]. There are many clustering algorithms. Some of them share the input set on a certain number of clusters, some other of them automatically select the number of clusters.

Algorithm DBSCAN (Density Based Spatial Clustering of Applications with Noise) is “closeness” algorithm for clusterization of spatial data with presence of noise with automatic selection of clusters number. It is based on the assumption that the density of points inside cluster is higher than outside of clusters. This algorithm allows to find clusters of arbitrary shape (see Fig. 1).

Fig. 1. Examples of clusters of arbitrary shape, producing flat breakdown [4]

The algorithm was proposed by Ester M., Kriegel H.-P and colleagues for solution of data partitioning (firstly spatial) on clusters of arbitrary shape [4, 5]. Most of algorithms create the clusters having form are close to spherical because minimizing the distance of objects from the center cluster. Authors of DBSCAN experimentally showed that their algorithm is able to identify clusters of different shapes, such as above in Fig. 1. The idea underlying the algorithm lies in the fact that within each cluster there is a typical density of points (objects) which have considerably higher than the density outside of the cluster and density in areas with noise have the lower density of each clusters.

In this paper at the decision of applied problems in the field of bioinformatics data processing have been proposed the inductive clustering algorithm with built-in algorithm of DBSCAN, in which decisions are approximate to global minimum that receive by successive run of DBSCAN operations.

2 Problem statement

The general statement of the cluster analysis problem in a broad sense is:

1/ the selection of the optimal clustering with

2/ simultaneous selection of the optimal ensemble of informative features within a given set of criteria, our problem can be formally represented as follows.

Suppose that given the set of $\{x_{0j},\ j=1,\dots,m\}$ – of target features. Let the space of attributes is $x_{ij} \in X$ $(i=1,\dots,n,\ \ j=1,\dots,m)$.

Thus, the total array of input data in our problem is:

$$\tilde{X} = \left(x_{0j} \vdots x_{ij} \in X\right),\ j=\overline{1,m}, i=\overline{1,n}\ . \qquad (1)$$

It is necessary:

1) to synthesize a subset $\{x_{\eta}^{*}\} = X^{*} \subset X$, $\eta = 1,\dots,n^{*}, n^{*} \le n$ of the above features better by a predetermined optimality criterion and which would allow:

2) to classify all objects on $k < m$, $k = 1,\dots,$ K , homogenous groups (clusters).

3 Solution of Problem

Some words about the optimality criterion of the regularized clusterization. It is known that application of inductive modeling of complex systems methodology [6] to obtain an optimal clustering requires the separation input set of objects $\omega_k \in \Omega$ to be clustering by at least on two disjointed subsets of Ω^A and Ω^B, wherein: $\Omega^A \cup \Omega^B = \Omega,\ \Omega^A \cap \Omega^B = \varnothing$.

Let on subsets of Ω^A and Ω^B by one of the chosen procedures of cluster-analysis obtained the clusterizations $s_t^A \in S^A$ and $s_t^B \in S^B$ with the same number of clusters $k_t{}^A = k_t{}^B = K_t$ (t – number of clustering corresponding to a subspace of features $X_t \subset X$, $k_t^{(\cdot)}$ – the number of clusters in the t-th clusterization) in the Euclidean subspace $X_t \subset X$ and suppose that for all K_t clusters from s_t^A and s_t^B their centers $\hat{m}_k^A$ and $\hat{m}_k^B$, $k=1,\dots,K_t$ calculated along the axis of the target feature x_0.

Then the optimality criterion of the regularized clusterization can be written in the simplest and more generally way as:

$$\rho^2(\hat{m}) = \sum_{k=1}^{K} (\hat{m}_k^A - \hat{m}_k^B)^2 \to \min\ . \qquad (2)$$

Criterion (2) requires that the sum of squared deviations between the centers of the clusters on the axis of the target feature installed on subsets Ω^A and Ω^B will be as minimal. Therefore, the such criterion in our case also be called as the criterion of *least inter-center deviations* (LICD), which is obviously a member of the class of *external consistency (of unbiasedness) criterion* in the GMDH, which is widely used in the inductive modeling procedures.

Fig. 2 shows the operating principle of LICD-criterion $\rho^2(\hat{m})$ at $n_k = 3$.

The inductive method of clustering with integrated genetic algorithm is shown in Fig. 3. The procedure of inductive algorithm brought in [7].

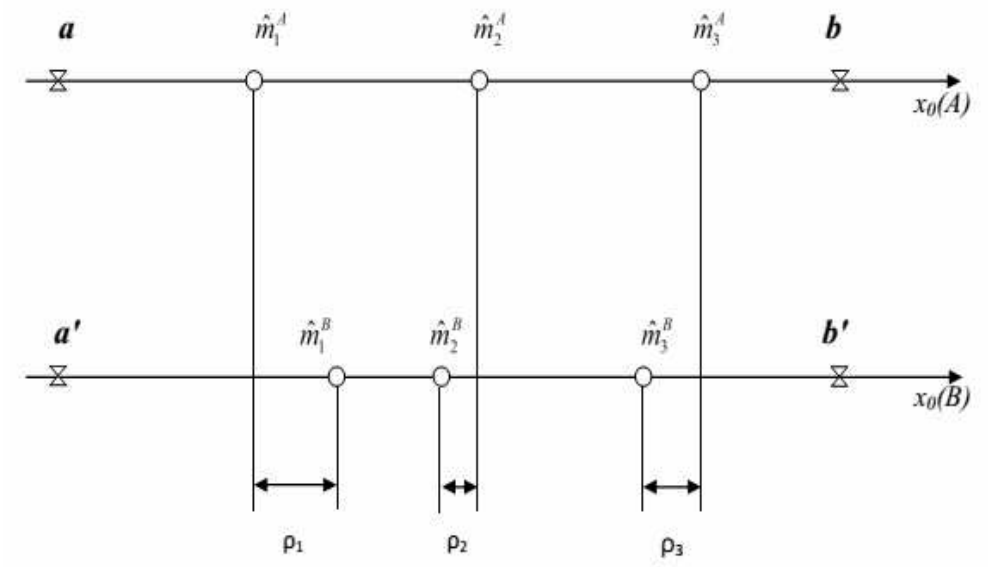

Fig. 2. To the operation principle of LICD (the unbiasedness of clustering: $\rho_\Sigma = |\rho_1| + |\rho_2| + |\rho_3|$)

Step 1. Selection or pre-synthesis target attribute x_0 to all $\omega_k \in \Omega$.

Step 2. Separation (1) into two parts A and B (Ω^A and Ω^B), according to claims of group method of data handling [6], [8]. The prepared general matrix $\tilde{X}$ of the input data would have a conditional form:

$$\tilde{X} = \left[\left(x_{0j} \vdots X \right)^A \vdots \left(x_{0j} \vdots X \right)^B \right], \quad j = 1,...,m^A = m^B, \; m^A + m^B = m. \tag{3}$$

Step 3. Setting up the clustering procedure DBSCAN. At this stage nature of $\tilde{X}$ plays an important role.

Step 4. Clustering of objects $\omega_k \in \Omega$ using the selected algorithm and configured independently on subsets Ω^A and Ω^B in space $\tilde{X}$ by one of the classic schemes of GMDH-algorithms with inductive increasing the number of features in their ensembles. The multiple-row inductive procedure of clustering, for example, would be like that.

1st row selection:

1.1) clustering of objects from subsets Ω^A and Ω^B for ensembles $\{x_i\}, i = 1,...,n$;

1.2) projecting centers clusters obtained per axis x_0;

1.3) for clustering, in which the condition $k_t^A = k_t^B = K_t$ (t - the current number of clustering; $k_t^{(\cdot)}$ - the number of clusters in the t-th clustering) is satisfied, the values of the optimality criterion of $\rho^2(\hat{m})$ are calculated;

2nd row selection:

2.1) Clustering objects on subsets Ω^A and Ω^B for ensembles $\{x_i, x_j\}$, $i, j = 1,...,n$, $i \neq j$;

2.2) the pp. (1.2) - (1.4) are performed and the best F ($F \leq n$) clusterizations S_f, $f = 1,...,F$ as well as the corresponding ensembles of features X_f, $f = 1,...,F$ are selected by system of criteria (2).

Third and subsequent rows selection:

3.1) clustering of objects on subsets Ω^A and Ω^B for ensembles $\{X_f, x_l\}$, $f = 1,...,F$, $l = 1,...,n$, provided that the feature with index l is not present in the already established ensembles X_f.

3.2) running part (2.2).

Stopping rule: inductive procedure is interrupted, provided that:

$$\rho^2(\hat{m})_s \leq \rho^2(\hat{m})_{s+1} \tag{4}$$

where s - is a number of selection in terms of GMDH. In this case, the value $k^{*(A)} = k^{*(B)} = K^*$, $K^* \leq m/2$ is fixed, and the corresponding subspace of informative features $\{x_l^*\} = X^*$, $l = 1,...,n^*$, $n^* \leq n$ selected as the best.

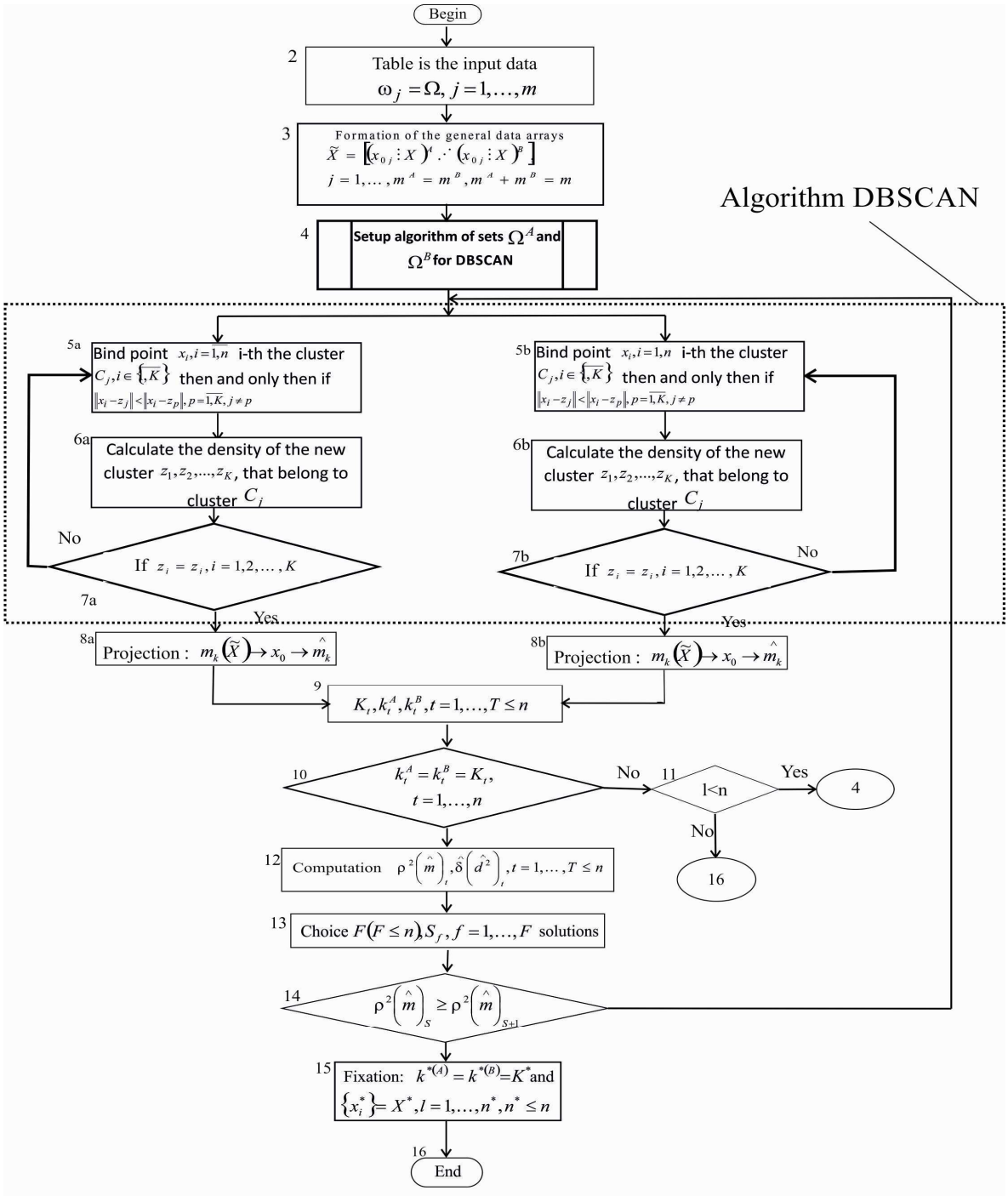

Fig.3. Scheme of the inductive method cluster analysis with the use of algorithm DBSCAN

DBSCAN algorithm boils down to this.

Input: the set of objects *S*, *Eps* and *MinPt*.
An object can be in one of three states:
1. Not marked.
2. Marked, which is not internal subject of any cluster.
3. Attributed to some cluster.

Step 1. Set the all elements of the set *S* flag "not marked." Assign current cluster C_j zero number. For noise set of points assign: Noise = 0 points.

Step 2. For each $s_i \in S$ such that flag (s_i) = "not marked" do:

Step 3. Flag (s_i): = "marked".

Step 4. $N_i = N_{Eps}(s_i) = \{q \in S | dist(s_i, q) \le Eps\}$.

Step 5. If $|s_i| < MinPt$,

then $Noise = Noise + \{s_i\}$.

In other case: the number of the next cluster $j = j + 1$;

EXPANDCLUSTER $(s_i, N_i, C_j, Eps, MinPt)$;

Output: a set of cluster $C = (C_j)$s.

EXPANDCLUSTER

Input: current object s_i, its eps -neighborhood N_i, the current cluster C_j and $Eps, MinPt$.

Step 1 $C_j = C_j + \{s_i\}$. ;

Step 2. For all points $s_k \in N_i$;

Step 3. If flag (s_k) = "no marked" then

Step 4. The flag (s_k) = "marked";

Step 5. $N_{ik} = N_{Eps}(s_k)$;

Step 6. If $|N_{ik}| \ge MinPt$ then $N_i = N_i + N_{ik}$,

Step 7. If $\nexists\ p : s_k \in C_p, p = \overline{1,(C)}\}$, than

$C_j = C_j + \{s_k\}$;

Output: cluster C_j.

4 Conclusion

The “closeness” clusterization algorithm of spatial data with presence of noise DBSCAN have been studied and implemented. The main advantage of this algorithm consists in the fact that it is suitable for the selection of arbitrary shape clusters. Computational complexity of algorithm *is* $O(n^2)$. Its main drawback is that exist the potential problem if the density of different clusters differs considerably. Are interested in using clustering algorithms that take into account the dynamic nature of the data set.

References

[1]. Duran B. Cluster analysis / B. Durand, P. Odell. — M.: Statistics, 1977. — 128 p. [In Russian].

[2]. Zagoruiko NG Applied methods of data analysis and knowledge / NG Zagoruiko. - Novosibirsk: Publishing House of the Institute of Mathematics, 1999. — 270P. [In Russian].

[3]. Lytvynenko V. Cluster analysis of data based on the modified immune network / V. Lytvynenko //USiM. — 2009. — №1. — Pp. 54-61, 85. [In Russian].

[4]. Ester M., Kriegel H.-P., and Xu X. 1995. A Database Interface for Clustering in Large Spatial Databases, Proc. 1st Int. Conf. on Knowledge Discovery and Data Mining, Montreal, Canada, 1995, AAAI Press, 1995.

[5]. M. Daszykowski, B. Walczak, D. L. Massart, Looking for Natural Patterns in Data. Part 1: Density Based Approach, Chemom. Intell. Lab. Syst. 56 (2001) 83-92

[6]. Ivachnenko A.G. The inductive method of self-organizing models of complex systems / A. Ivachnenko.—Kyiv: Naukova Dumka.—1982. — 296 p. [In Russian].

[7]. Osypenko V.V. The inductive algorithm of cluster analysis in the toolbox of system information-analytical research / V.V. Osypenko // USiM. — 2013. — № 2. — Pp. 59-64. [In Russian].

[8]. Ivachnenko A.G. The use of multi-alternative pattern recognition algorithms and group method of data handling for processing of expert assessment in global capital investment projects / A.G. Ivachnenko, E.A. Savchenko, G.A. Ivachnenko, T. Gergely // Cybernetics and calc. technique. MY. 133. — 2001. — C. 3-7. [In Russian].