Research and Development of Methods and Algorithms Non-hierarchical Clustering

Yuri Stekh, Mykhaylo Lobur, Vitalij Artsibasov

Abstract – **Researched and developed the methods and non-hierarchical clustering algorithms for determining the optimal initial number of clusters without any background information on the location of the clusters. The methods and algorithms are researched in the famous test set Iris.**

Index Terms—**Non-hierarchical clastering, initial number of clusters, clustering algorithms library, clustering algorithms testing.**

I. Introduction

THE issue of clustering is one of the fundamental tasks in Data Mining, Web Mining, Text Mining, machine learning [2,5,6]. The main task of clustering is to partition a given set of images into classes (clusters) that allow you to explore the similarities and differences between the images in the clusters and make sound conclusions about the images that belong to certain clusters. In clustering process is not any information about the predefined classes. And therefore the process of clustering refers to the problems of unsupervised learning. The result of clustering depends on several factors that determine of which is a method and a clustering algorithm, the initial parameters of clustering algorithm. The main steps in the process of clustering are shown in Fig. 1.

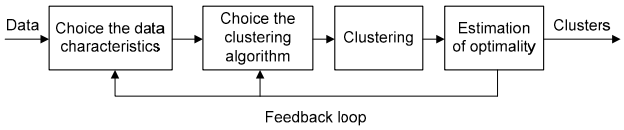


Fig. 1. The main steps of the clustering process

Manuscript received April 20, 2011. This work was supported by Department of Computer-Aided Design Systems (Lviv Polytechnic National University) and the Faculty of Materials Science and Engineering (Warsaw University of Technology).

Yuri Stekh is with the Lviv Polytechnic National University, Ukraine (e-mail: stekh@polynet.lviv.ua).

Mykhajlo Lobur is with the Lviv Polytechnic National University, Ukraine (e-mail: mlobur@polynet.lviv.ua).

Vitalij Artsibasov is with the Lviv Polytechnic National University, Ukraine (e-mail: free_2@list.ru).

At the stage of feature selection of images the feature vectors of images are generated that best reflect the properties of objects that are clustered. At the selection stage of clustering algorithm will be selected one of the algorithms contained in the library of algorithms for clustering. This stage involves the selection of a similarity measure of images in the cluster and the clustering criteria. Similarity measure is set as the basis of the rules for inclusion of the image to a particular cluster. Clustering criterion determines the stop of clustering algorithm. At the stage of clustering is clustering a given set of images with chosen algorithm for the chosen degree of similarity criteria and clustering.

In the evaluation of clustering results assesses the optimality of the partition a given set of images into clusters. At this stage are used some precise and / or approximate optimality criteria. Therefore, finding an optimal partitioning into clusters requires the construction and study of ensemble methods and algorithms for clustering and application of several criteria for optimality of the partition into clusters.

II. Formal Problem Statement

Let $D = \{\overline{X}_1, \overline{X}_2, ..., \overline{X}_n\}$ - a set of n images, each of which has d attributes. Let $L = \{A_1, A_2, ..., A_m\}$ - a set of algorithms for clustering. Each clustering algorithm A_i generates a partition for a given set D into clusters $P^{(i)} = \{S_1^{(i)}, S_2^{(i)}, ...S_m^{(i)}\}$. Let $S = \bigcup_{i=1}^{m} P^{(i)}$ is set of all clusters, which received with a set of algorithms L. The aim of the ensemble methods and algorithms is to obtain the optimal partition into clusters corresponding to the optimal maximum or minimum value of the optimality criteria $T = \{C_1, C_2, ...C_k\}$. The elements of T must satisfy the following properties:

- Each cluster must have at least one image $C_i \neq \varnothing \quad \forall i \in \{1,2,...,k\}$
- Each image must belong to at least one cluster $C_i \cap C_j = \varnothing \quad \forall i \neq j, i, j = \{1,2,...,k\}$
- All images must be separated by cluster $\bigcup_{i=1}^{k} C_i = P$

Thus, the clustering problem is reduced to an optimization problem that requires research and development of ensemble methods and algorithms, and study certain criterion functions.

III. Analysis of Know Solutions to the Problem

Well-known clustering algorithms can be divided into hierarchical and nonhierarchical [2,6]. In hierarchical clustering algorithm refuse to determine the number of clusters. Instead, it builds a tree of nested clusters (dendrogram). Problems such algorithms are well known: the choice of measure of the closeness of clusters, the problem of the inverted index in dendrogram, the inflexibility of the hierarchical clustering.

In non-hierarchical clustering algorithms, the nature of their work and conditions of the stop must be set in advance using the input parameters of the algorithm. The most important of these is the number of desired clusters.

IV. Development of Methods and Algorithm For Non-Hierarchical Clustering

Solving the problem of optimal choice of the initial number of clusters is proposed to resolve through the development of ensemble methods and algorithms for non-hierarchical clustering. The overall design and the use of algorithms are shown in Fig. 2.

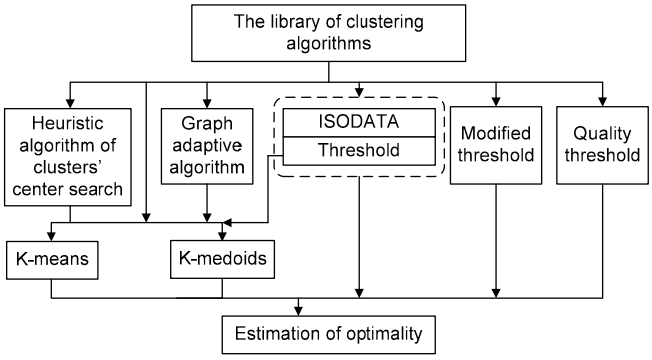


Fig. 2. The overall design and the use of algorithms

The main feature of the developed non-hierarchical clustering algorithms library is that the problem of choosing the number of clusters is solved by using two algorithms: a heuristic search algorithm of the cluster centers and cluster centers of the search algorithm using a neural network [7]. Heuristic search algorithm of the cluster centers (HSACC) is designed to find cluster centers in a given set of images without any background information on the location of the cluster centers. The algorithm is implemented in two versions. The first modification finds the most distant point of each cluster centers. After finding the centers of the clusters remaining images are distributed in clusters on the criterion of minimum Euclidean distance to cluster centers.

Let $\overline{X} = \{\overline{x}_1, \overline{x}_2, ..., \overline{x}_n\}$ - a set of points of images, $\overline{Z} = \{\overline{z}_1, \overline{z}_2, ..., \overline{z}_n\}$ - the desired cluster centers.

Step 1. Select randomly the center point of the first cluster k.

Step 2. $\overline{z}_1 = \overline{x}_k, l = n - 1$

Step 3. Compute the Euclidean distance from the remaining points of the set of images to the center of the first cluster

$$D_{i1} = \| \overline{x}_i - \overline{z}_1 \|, i = 1(1)l$$

Step 4. $K_i^{(1)} = \max_i \{D_{i1}\}, L_1 = K_i^{(1)}, p = i$

Step 5. Select the center of the second cluster $\overline{z}_2 = \overline{x}_p, l = l - 1$

Step 6. Compute the Euclidean distance from the remaining points of the set of images to the center of the first and second clusters

$$D_{i1} = \| \overline{x}_i - \overline{z}_1 \|, D_{i2} = \| \overline{x}_i - \overline{z}_2 \|, i = 1(1)l$$

Step 7. $A_i = \min_i \{D_{i1}, D_{i2}\}$

Step 8. $K_i^{(2)} = \max_i \{D_{i1}, D_{i2}\}, L_2 = K_i^{(2)}, p = i$

Step 9. If $L_2 > S \cdot L_1$ then $\overline{z}_3 = \overline{x}_p, l = l - 1$, else STOP.

Step 10. Compute $L_{c.a.} = \frac{L_1 + L_2}{2}$.

Step 11. Compute the Euclidean distance from the remaining points of the set of images to the center of the first, second and third clusters: $D_{i1} = \| \overline{x}_i - \overline{z}_1 \|$ $D_{i2} = \| \overline{x}_i - \overline{z}_2 \|$, $D_{i3} = \| \overline{x}_i - \overline{z}_3 \|, i = 1(1)l$.

Step 12. Compute $A_i = \min_i \{D_{i1}, D_{i2}, D_{i3}\}, i = 1(1)l$.

Step 13. Compute $K_i^{(3)} = \max_i \{A_i\}, L_3 = K_i^{(3)}, p = i$.

Step 14. If $L_3 > S \cdot L_{c.a.}$ then $\overline{z}_4 = \overline{x}_p, l = l - 1$, else STOP;

$\overline{x}_i \in A_k$ if $\| \overline{x}_i - \overline{z}_k \| < \| \overline{x}_i - \overline{z}_r \|, r = 1(1)l, m \neq k$

The parameter S is chosen within $S \in (0,1)$.

Such a construction algorithm finds the most distant centers of the clusters, does not always lead to optimal outcome. Therefore the second modification was originally defined by the geometric center point set of images.

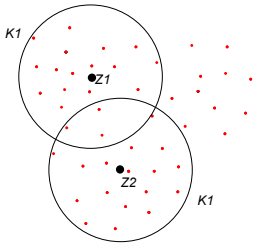
$$\overline{x}_c = \frac{1}{|D|} \sum_{i=1}^{n} \overline{x}_i \qquad (1)$$

where $|D|$ - cardinality of point set images.

The point of the first cluster center is chosen as the most distant point on $\overline{x}_c$. In many cases, this allows us to determine more optimal centers of the clusters. An alternative method of finding the initial cluster centers in the developed ensemble methods and algorithms is proposed in [7] algorithm for finding the cluster centers by graph algorithm. In this algorithm, investigated a set of

points be represented by a full connected undirected graph , where each image is associated with node. Thus the investigated set of points in the images of d-dimensional space turn into a complete undirected weighted graph.

The algorithm works in such a way that the nodes, that are on the border of the cluster region, transmit their activity to the nodes, located within the regions of clusters. The learning process of the complete undirected weighted graph converges to such a result, when in each cluster region is only one active neuron - the center of the cluster. This approach to finding the cluster centers can continue to use well-known algorithms for k-means and k-medoids [4,6], without specifying the initial parameters for them. Library of algorithms includes the well known threshold algorithm [4]. The main drawback of this algorithm is that it requires setting a threshold T and the output of this algorithm depends on the choice of starting point – the first cluster center. As a result, the algorithm can form a series of clusters that give the section of clustering regions and require additional heuristics to determine to which cluster belongs the point of images. The library including our modified threshold algorithm (Fig.3).



$D_{i1} = \|\bar{x}_i - \bar{z}_1\|$

$D_{i1} < T \rightarrow \bar{x}_i \in K_1$

$\bar{z}_2 = \frac{1}{n}\sum_{i=1}^{n} \bar{x}_i$

$\|\bar{z}_{i+1} - z_i\| \le \varepsilon$

Ai – Datapoints set in the cluster Ki

Bo – Initial set of datapoins

$B_{i+1} = B_i \setminus A_i$

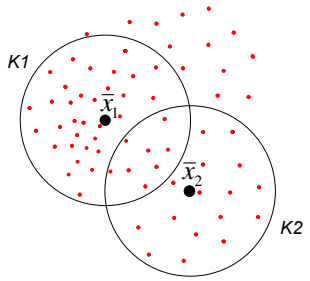
$B_{i+1} = \varnothing \rightarrow STOP$

Fig. 3. The modified threshold algorithm

This algorithm is a combination of classical threshold algorithm and k-means. Feature designed combination is that, unlike the threshold algorithm, in each area of clustering first calculated the geometric center of the set of images and the set of images of this area later excluded from consideration as the next cluster. The clustering process continues until we obtain an empty set of image points.

It is developed improved quality threshold algorithm. This algorithm represents an improvement in the modified threshold algorithm (Fig. 4).

In this algorithm first the clusters for each image of a given set is computed with steps of a modified threshold.algorithm. The number of points of the images in each field are computed. As the first cluster are taking the region with the highest number of points of images. This set of points is eliminated from further consideration. This iterative process continues until we obtain an empty set of points in the images. This algorithm has the highest complexity among all algorithms library. The developed algorithms and combinations of algorithms were tested on a known test set Iris [5]. The results of the test are reported in Table 1.



Bo – Initial set of datapoins

$|B_0| = m$

$K_1 = n_1, K_2 = n_2, ..., K_m = n_m$

$g_i = \max_i (n_1, n_2, ..., n_m)$

Ki – First cluster

$B_1 = B_0 \setminus K_i$

.....

$B_{p+1} = B_p \setminus K_i$

$B_{p+1} = \varnothing \rightarrow STOP$

Fig. 4. The improved quality threshold algorithm

TABLE I
THE RESULTS OF THE ALGORITHMS WORK TEST

Distance function \ Results	Algo-rithm	Percent of wrong cluste-red points (%)	Duration of algorithm work (ms)	Mean size of cluster	Mean distance between clusters	Mean distance between cluster centers
Euclidian	K-means	14,67	6419	0,9212	3,473	3,366
	HSACC+ K-maens	16	3510	0,9183	3,487	3,4
	HSACC+ K-medoids	16	2787	0,9183	3,487	3,433
Square Euclidean	K-means	12	2694	1,099	12,33	11,133
	HSACC+ K-means	20	3026	1,068	14,6	13,766
	HSACC+ K-medoids	20	2797	1,068	14,6	13,666
Powered (p=4, r=2)	K-means	10,67	3273	0,7582	9,159	8,433
	HSACC+ K-means	20	4072	0,7431	10,77	10
	HSACC+ K-medoids	17,33	2276	0,7425	10,6	10,466

V. CONCLUSION

Devepoled methods and algorithms for non-hierarchical clustering. Developed a software library that allows to find the optimal partition into clusters using computing clusters with different algorithms with different initial parameters and by computing the criterial functions such as the average

intercluster distance, the average size of the cluster and the average distance between the centers of the clusters.

REFERENCES

[1] Barseghyan A.A, Kupriyanov M.S, Stepanenko V.V., Kholod I.I. Models and methods of data analysis: OLAP and Data Mining.- Petersburg: BHV - Petersburg, 2004. 236 p. (in Russian)

[2] Barseghyan A.A., Kupriyanov M.S, Stepanenko V.V., Kholod I.I. Technology of data analysis: Data Mining, Visual Mining, Text Mining, OLAP. SPb.: BHV - Petersburg, 2007. 384 p. (in Russian)

[3] Kim Dzh.-O., Myuller C.W., Klekka W.R. Factorial, discriminant and cluster analysis. Per. from English. Moscow: Finances and Statistics, 1989. 215 p. (in Russian)

[4] Tu J., Gonzales R. Principles of pattern recognition, 1978. 411 p.

[5] Classification and cluster / J. Van Rayzin. New York. 389 p.

[6] A. J. Jain, M. N. Murty, P.J. Flynn "Data clustering: a review" *ACM Computing Surveys, v..31, pp. 264-323, 1999.*

[7] I. Farmaga, P. Shmigelskyi, P. Spiewak, L. Ciupinski. Evaluation of Computational Complexity of Finite Element Analysis // IEEE CADSM'2011. – Polyana, 2011. – PP. 213 - 214.

[8] Stekh Y., M.E. Faisal Sardis, Lobur M.V., Kernytskyy A.B. Algorithm for finding the optimal number of clusters // Bulletin of the National University "Lviv Polytechnic". № 651. P.129-132. 2009 (in Ukrainian)

Dr. Stekh Yuri, PhD, Assistant Professor of CAD department of Lviv Polytechnic National University. Research interests: databases and artificial intelligence methods for CAD tools for microelectronic devices design, Data Mining.

Prof. Mykhaylo Lobur, DSc, PhD, Academitian of Academy of Science in applied radioelectronics, Head of CAD department of Lviv Polytechnic National University, Director of LRERI. Research interests: design of CAD tools for microelectronic devices design.

Vitalij Artsibasov Master of Computer Science. Research interests: artificial intelligence, Data Mining.