https://doi.org/10.52326/ic-ecco.2022/CS.07

Pseudo Genetic Algorithm of Clustering For Linear and Ellipsoidal Clusters

Valerii Fratavchan [1], ORCID: 0000-0001-9347-9855
Tonia Fratavchan [1], ORCID: 0000-0003-1076-0794
Victor Ababii [2], ORCID: 0000-0002-0769-8144
[1] Yuriy Fedkovych Chernivtsi National University, Chernivtsi, Ukraine, v.fratavchan@chnu.edu.ua
[2] Technical University of Moldova, Chisinau, Republic of Moldova

***Abstract*—this article considers the method of clustering in the problems of pattern recognition when studying with a teacher in the case of n-dimensional numerical features. Clusters of linear and ellipsoidal forms that are optimal in the number of errors are created by the method of pseudo genetic algorithm. The pseudo genetic algorithm has a simplified procedure for performing mutation and crossover operations.**

Keywords – cluster; clustering; linear cluster; ellipsoidal cluster; genetic algorithm; mutation; crossover; discriminant function.

I. Introduction

Let’s define a set of objects belonging to K classes. Each object is described by an n-dimensional vector of numerical features $X=(x_0,x_1,...,x_{n-1})$. Each feature is a random variable. Objects of the same class are distributed in the feature space according to the near normal law. Each class in the feature space takes a particular area. Moreover, class areas can intersect. The problem is that it is necessary to divide the feature space for each class into areas that do not intersect (clusters) in such a way that the number of objects, which are not localized in the cluster of "their" class, is minimal [1]. The clustering process takes place according to the method of "learning with the teacher". Also, a learning sequence of images, which consists of representative sets of samples of each class $T=\left(T^{(1)},T^{(2)},...,T^{(K)}\right)$, has been created. What is created in the feature space is an object area that consists of clusters $C=\left(C^{(1)},C^{(2)},...,C^{(K)}\right)$. Consequently, each item of the training sequence must belong to only one cluster $X^{(i)}\in C^{(j)},(X^{(i)}\notin C^{(k)},j\notin k)$. Considering everything said above, it is desirable that the clusters in the feature space have a convex shape. As well as that, the classification procedure has to be quite simple. Such requirements are met by clusters of linear and ellipsoidal configurations [2].

II. Construction of Linear Clusters

Linear clusters have a convex polygonal shape. There is a hyperplane that optimally separates one cluster from another in an n-dimensional space for any two non-matching clusters (Figure 1).

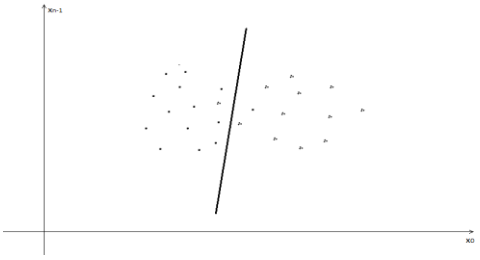

Figure 1. Hyperplane between clusters.

In the case of K classes, it is necessary to find $K-1$ such hyperplanes for each cluster, in the total of $K\left(K-1\right)/2$, since for each pair the hyperplanes coincide (Figure 2). Polygonal clusters can have an open or closed shape.

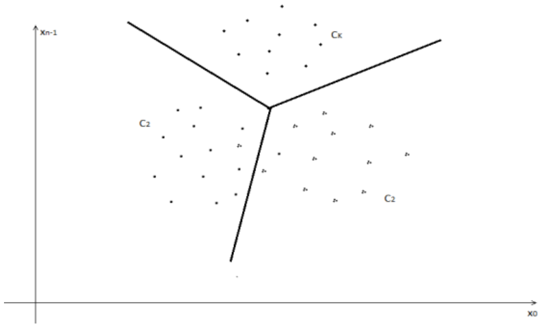

Figure 2. Polygonal clusters.

Mathematically, each cluster is defined by a system of inequalities:

$$\begin{cases} a_{1,0}^{(k)}x_0 + a_{1,1}^{(k)}x_1 + \ldots + a_{1,n-1}^{(k)}x_{n-1} + b_1^{(k)} \le 0 \\ a_{2,0}^{(k)}x_0 + a_{1,1}^{(k)}x_1 + \ldots + a_{2,n-1}^{(k)}x_{n-1} + b_2^{(k)} \le 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{K-1,0}^{(k)}x_0 + a_{1,1}^{(k)}x_1 + \ldots + a_{K-1,n-1}^{(k)}x_{n-1} + b_{K-1}^{(k)} \le 0 \\ k = 1,\ldots,K. \end{cases}$$

Thus, the set of clusters is determined by the set of weighting coefficients:

$$W = \begin{Bmatrix} a_{j,i}^{(k)}, (k = 1..K,\ j = 1..K-1, i = 0..n-1) \\ b_j^{(k)}, (k = 1..K,\ j = 1..K-1) \end{Bmatrix} \quad (1)$$

III. Construction of Ellipsoidal Clusters

Among the pattern recognition methods, there is a quite popular method of discriminant functions [4] such as

$$g_k(X) = \begin{cases} v \ge 0,\ X \in C^{(k)} \\ v < 0,\ X \notin C^{(k)} \\ k = 1..K \end{cases}.$$

In order to construct ellipsoidal clusters, it is suggested to use a discriminant function of the quadratic (parabolic) type (Figure 3).

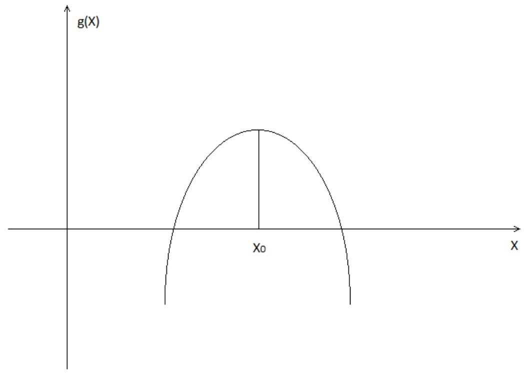

Figure 3. Parabolic discriminant function.

For the n-dimensional case, such a function will be set by the formula below:

$$g_k(X) = -a_0^{(k)}(x_0 - c_0^{(k)})^2 - a_1^{(k)}(x_1 - c_1^{(k)})^2 - \ldots \\ - a_{n-1}^{(k)}(x_{n-1} - c_{n-1}^{(k)})^2 + b_k$$

In n+1-dimensional space, these functions form paraboloids, which form ellipsoids in the n-dimensional feature space (Figure 4).

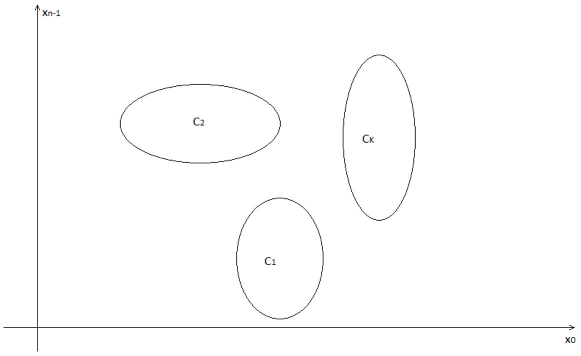

Figure 4. Ellipsoidal clusters.

Similarly to the case of linear clusters, elliptical clusters are determined by the set of coefficients of the discriminant functions and by the coordinates of the ellipsoid centers:

$$W = \begin{Bmatrix} a_i^{(k)}, k = 1..K, i = 0..n-1; \\ b_k, k = 1..K; \\ c_i^{(k)}, k = 1..K, i = 0..n-1 \end{Bmatrix}. \quad (2)$$

It can be noted that for each elliptic cluster it is necessary to define a $2n+1$ parameter that in case of dealing with numerous classes requires significantly less computing resources than for linear clusters. However, at the same time, the efficiency of clustering with ellipsoidal shapes can be inferior to linear clusters.

IV. Applications of Pseudo Genetic Algorithm for Building Optimal Clusters

Cluster construction procedures are the same for linear and elliptical clusters and are implemented as optimization problems [3]:

$$\begin{aligned} &E(W) \to \min, \\ &W \in D. \end{aligned} \quad (3)$$

Where: $E(W)$ – is the function that estimates the number of errors for the training set;

W – aggregate vector of cluster parameters according to formulas (1) or (2);

D – a domain of definition of cluster parameters.

The optimization problem is solved by a pseudo genetic algorithm. It has a simplified system of mutation and crossover genetic operations. In the classical genetic algorithm, these operations are performed on the binary code of "chromosomes", thus the values of specific bits change. In the pseudo genetic version, operations are performed on the components of the parameter vector.

For **mutation**, we randomly select a specimen of the current population and the index of some component of the vector of cluster parameters, after this the component is given a new random value from the domain of definition:

$$\{w_1,..., w_m,..., w_s\} \rightarrow \{w_1,..., u_m,..., w_s\} ,$$

Where: $m-$ is the index of the component selected for mutation:

s - the size of the parameter vector;

$u_m = random(d, d \in D)$.

For the **crossover** operation, we randomly select two "parent" instances of parameter vectors from the current population and the crossover point index. Eventually, the information is being exchanged relative to this point:

$$\begin{pmatrix} \{w_1,..., w_m, w_{m+1},..., w_s\}, \\ \{u_1,..., u_m, u_{m+1},..., u_s\} \end{pmatrix} \rightarrow .$$

$$\begin{pmatrix} \{w_1,..., w_m, u_{m+1},..., u_s\}, \\ \{u_1,..., u_m, w_{m+1},..., w_s\} \end{pmatrix}$$

The general procedure of the genetic algorithm corresponds to the classical scheme and consists of the formation of the initial population of various cluster system configurations, and the population regeneration process with the usage of mutation, crossover and selection operations. The stages of regeneration continue until the value of the evaluation function is stable or until the desired qualitative assessment of the classification is achieved when processing the training set samples.

CONCLUSIONS

The purpose of the paper is to suggest algorithms for clustering a multidimensional feature space in pattern recognition tasks in the learning mode with a teacher. Summing up the results, it can be concluded that algorithms are built on relatively simple mathematical models and do not require powerful system resources. This paper has clearly shown that algorithms are quite effective if the selected features allow localizing clusters in the feature space with areas of convex shape. However, the algorithms are not effective in large estimates of the mutual intersection of the internal parts of the cluster linear shells. To implement clustering procedures, it is necessary to a priori create representative training sequences, in which the normal laws of image distribution in the cluster become sufficiently noticeable. These features require preliminary statistical analysis of training sets.

The obtained data indicate that these algorithms are recommended for use in cybernetic stand-alone systems, where the application of highly effective library intellectual tools is obstructed.

REFERENCES

[1] Фратавчан В.Г., Фратавчан Т.М., Сугак І.С. Концептуальна схема побудови системи розпізнавання у n-вимірному просторі ймовірнісних ознак. Проблеми інформатики і комп’ютерної техніки (ПІКТ-2020) : Праці IX міжнар. наук.-практ. конф., м. Чернівці, 28-31 жовт. 2020 р. Чернівці : Черн. нац. ун-т, 2020. С. 119-120.

[2] V. FRATAVCHAN, T. FRATAVCHAN, One Pattern Recognition Method for Complex Geometric Clusters Configuration//Proceedings of the 14th International Conference on Development and Application Systems, DAS 2018 (24-26, May 2018, Suceava - Romania), pp.200-203.

[3] V. Fratavchan, The Application of Genetic Algorithm for Training “Without a Teacher”, //Proceedings of the 10th International Conference on Development and Application Systems, DAS 2010 (27-29 May 2010, Suceava - Romania), pp.105-107.

[4] Зайченко Ю.П. Основи проєктування інтелектуальних систем. Навчальний посібник. – К.:Видавничий дім «Слово», 2004. – 352с.