O NOUĂ ŞI ROBUSTĂ TEHNICĂ DE CLUSTERING DESTINATĂ PROGRAMELOR DE PROIECTARE OPTIMALĂ

Daniela JUCAN, Lucian TUDOSE
Universitatea Tehnică Cluj-Napoca

In this paper, a new simple and robust clustering technique addressed to optimal design is presented. This technique is based on the assumption that any couple formed by a set of points and its envelope, in the n-dimensional space, is characterized by a unique real number, named $t_{\Omega m}$-threshold that is able to describe the state of uniform cover. The algorithm was implemented in our new software (*GetCluster v. 1.0*).

1. NECESITATEA TEHNICILOR DE CLUSTERING IN PROIECTAREA OPTIMALĂ

În proiectarea optimală, Algoritmii Genetici reprezintă, poate, cele mai utilizate metode de optimizare. O problemă importantă este aceea a gestionării funcțiilor obiectiv multimodale. Autorii lucrării consideră că se poate dezvolta o nouă categorie de algoritmi de optimizare care să includă module de clustering. Aceste module vor fi necesare în următoarea modificare a Algoritmului Genetic:

La un moment dat evoluția populației se opreşte şi se identifică aglomerările, grupurile (clusterele) din spațiul variabilelor (genelor) cu speranța ca aceste clustere, lăsate ulterior să evolueze ca populații independente, să conducă, fiecare, la identificarea unui alt punct de extrem local al funcției obiectiv (multimodală).

Este evident că un astfel de algoritm de clustering nu va fi necesar să fie extrem de precis, ci va trebui să aibă alte calități, şi anume:

- să identifice rapid numărul de clustere existent în populație;
- să fie cât mai simplu;
- să fie foarte rapid;
- să fie robust.

Autorii lucrării propun aici o nouă tehnică de clustering, ce a dat rezultate foarte bune în implementările ei în programele proprii de optimizare cu Algoritmi Genetici.

De remarcat că această tehnică a fost implementată cu succes şi într-un modul de *pattern recognition* necesar funcționării unui robot industrial.

2. NOUA TEHNICĂ DE CLUSTERING

2.1 Axiome şi definiții

Se consideră o mulțime $\Omega_m \subset \mathbb{R}^n$, adică o listă neordonată ce cuprinde coordonatele, în spațiul cu *n* dimensiuni, a celor *m* puncte ce urmează a fi supuse clustering-ului.

$$\Omega_m = \left\{ M_1\left(x_1^1, x_2^1, \ldots, x_n^1\right), M_2\left(x_1^2, x_2^2, \ldots, x_n^2\right), \ldots, M_m\left(x_1^m, x_2^m, \ldots, x_n^m\right) \right\} \tag{1}$$

Reamintim că nu se cunoaşte *a priori* numărul de clustere existent în mulțime şi că această sarcină revine algoritmului propus. Pentru înțelegerea tehnicilor noi de *clustering* propuse, se propun aici câteva axiome şi definiții ale noțiunilor importante:

Definiția 1. *Pentru fiecare mulțime Ω_m se defineşte n-paralelipipedul cuprinzător ca fiind mulțimea:*

$$S_{\Omega m}=\left\{\left(x_1,x_2,\ldots,x_n\right)\ \mid x_1^{\min}\le x_1\le x_1^{\max},\ x_2^{\min}\le x_2\le x_2^{\max},\ldots,x_n^{\min}\le x_n\le x_n^{\max}\right\} \tag{2}$$

unde

$$x_i^{\min}=\min\left\{x_i^1,x_i^2,\ldots,x_i^m\right\},\quad (\forall)i\in\{1,2,\ldots,n\}$$

$$x_i^{\max}=\max\left\{x_i^1,x_i^2,\ldots,x_i^m\right\},\quad (\forall)i\in\{1,2,\ldots,n\}$$

Axioma 1. *Pentru fiecare cuplu $\left(\Omega_m,S_{\Omega m}\right)$ există o unică valoare prag (threshold) $t_{\Omega m}>0$ ce descrie situația in care mulțimea Ω_m acoperă „uniform” n-paralelipipedul $S_{\Omega m}$.*

Axioma 2. *Mulțimea Ω_m acoperă „uniform” n-paralelipipedul $S_{\Omega m}$ dacă şi numai dacă distanța euclidiană dintre oricare două puncte ale mulțimii Ω_m este egală cu pragul $t_{\Omega m}$:*

$$d\left(M_i,M_j\right)=\sqrt{\left(x_1^i-x_1^j\right)^2+\left(x_2^i-x_2^j\right)^2+\ldots+\left(x_n^i-x_n^j\right)^2}=t_{\Omega m},\ (\forall)i,j\in\{1,2,\ldots,m\},\ i\ne j \tag{3}$$

Definiția 2. *Fie $M_i\in\Omega_m$ un punct oarecare. Punctul $M_j\in\Omega,\ j\ne i$ se numeşte $t_{\Omega m}$-vecin (neighbour) punctului M_i dacă:*

$$d\left(M_i,M_j\right)=\sqrt{\left(x_1^i-x_1^j\right)^2+\left(x_2^i-x_2^j\right)^2+\ldots+\left(x_n^i-x_n^j\right)^2}\le t_{\Omega m},\ \ i\ne j \tag{4}$$

Definiția 3. *O submulțime $C_{\Omega m}\in\Omega_m$ este un $t_{\Omega m}$- grup (cluster) dacă oricare două puncte $M_{i,j}\in C_{\Omega m},\ j\ne i$ sunt $t_{\Omega m}$- vecine (neighbours):*

$$d\left(M_i,M_j\right)=\sqrt{\left(x_1^i-x_1^j\right)^2+\left(x_2^i-x_2^j\right)^2+\ldots+\left(x_n^i-x_n^j\right)^2}\le t_{\Omega m},(\forall)M_{i,j}\in C_{\Omega m},\ j\ne i \tag{5}$$

Definiția 4. *Numărul de clustere al mulțimii Ω_m este numărul de $t_{\Omega m}$-grupuri (clusters) în care se grupa mulțimea respectivă.*

2.2 Calculul valoarii de prag $t_{\Omega m}$

Este evident că identificarea corectă a valorii de prag $t_{\Omega m}$ este esențială pentru construirea unui algoritm eficient pentru clustering-ul unei mulțimi. Autorii articolului propun următoarea definiție pentru $t_{\Omega m}$:

Definiția 5. *Valoarea de prag $t_{\Omega m}$ ataşată cuplului $\left(\Omega_m,S_{\Omega m}\right)$ se defineşte ca fiind lungimea laturii simplexului S_{Xn} care are volumul de m ori mai mic decât volumul n-paralelipipedului $S_{\Omega m}$:*

$$t_{\Omega m}=l\left(V\left(S_{Xn}\right)\right)=l\left(\frac{V\left(S_{\Omega m}\right)}{m}\right) \tag{6}$$

Pentru exemplificare, se vor trata aici numai cazurile 2D şi 3D.

2.2.1. Valoarea de prag $t_{\Omega m}$ în cazul 2D

Se consideră o mulțime $\Omega_m\subset\mathbb{R}^2$ ce cuprinde *m* puncte din plan ce urmează a fi supuse clustering-ului.

$$\Omega_m=\left\{M_1\left(x_1^1,x_2^1\right),M_2\left(x_1^2,x_2^2\right),\ldots,M_m\left(x_1^m,x_2^m\right)\right\} \tag{7}$$

Dreptunghiul cuprinzător (fig.1) va fi definit ca mulțimea:

$$S_{\Omega m}=\left\{\left(x_1,x_2\right)\ \mid x_1^{\min}\le x_i\le x_1^{\max},\ x_2^{\min}\le x_2\le x_2^{\max}\right\} \tag{8}$$