Clustering multi-objectifs basée sur l'algorithme d'essaim de salpedia bêta-distribués (Multi-Objectif Beta Salp Swarm Algorithm MO-β-SSA)

Yassmine Soussi, Nizar Rokbani, Mohamed Moncef Ben Khelifa, Ali Wali, Thanh Phuong Nguyen

Clustering multi-objectifs basée sur l'algorithme d'essaim de salpedia bêta-distribués
(Multi-Objectif Beta Salp Swarm Algorithm MO-β-SSA)

Yassmine Soussi [1-2-3-4], Nizar Rokbani[2] , Mohamed Moncef Ben Khelifa [4] , Ali Wali[2] , Nguyen Thanh Phuong [3]

[1] Université de Sousse, ISITCom, 4011, Hammem Sousse, Tunisie
[2] REGIM-Lab : REsearch Groups in Intelligent Machines, Université de Sfax, et Université de Sousse, Sousse, Tunisie.
[3] LIS: Laboratoire d'informatique et des systèmes, Université de Toulon, France.
[4] IAPS: Impact de l'activité physique sur la santé, Université de Toulon, France

soussi.yassmine@ieee.org

Résumé

*Les algorithmes d'optimisation du clustering multi-objectifs (**MOCO** algorithms), sont généralement utilisés pour prendre en compte divers critères de validité de cluster (**CVI**). Dans cet article, l'algorithme d'essaim de salpedia bêta-distribués, (**β-SSA**), a été utilisé dans sa forme multi objective (**MO-β-SSA**) et a été étudié pour les problèmes de MOCO. MO-β-SSA a été testé et comparé à des techniques de clustering connexes à travers un scénario utilisant trois CVI comme fonctions objectives (**OF**) : l'I-index, le Con-index et le Sym-index. Les performances de MO-β-SSA ont été fournies et comparées aux contributions associées en utilisant des bases de données (BD) standards.*

Mots Clés

Algorithme d'essaim des salpedia (**SSA)**, Agorithme d'essaim de salpedia bêta-distribués (**β-SSA)**, clustering, optimisation multi-objectif d'un clustering (**MOCO)**, indices de validité des clusters (**CVI)**.

Abstract

*Multi-Objective Clustering Optimization (**MOCO**) algorithms are generally used to take into account various cluster validity indices (**CVI**). In this article, the Salp Swarm Algorithm using a Beta-distribution (Beta Salp Swarm Algorithm **β-SSA**) was used in its multi-objective form (**MO-β-SSA**) and investigated for MOCO problems. MO-β-SSA was tested and compared to related clustering techniques through a scenario that uses three CVI as Objective Functions (OF): I-index, Con-index and Sym-index. MO-β-SSA performances were provided and compared to the associated contributions using standard databases.*

Keywords

Salp Swarm Algorithm **SSA**, Beta Salp Swarm Algorithm **β-SSA**, clustering, Multi-Objective Clustering Optimisation **MOCO**, Cluster Validity Indices **CVI.**

1 Introduction

Le regroupement ou clustering consiste à diviser une base de données (BD) en k clusters et ceci sur la base des similarités des objets appartenant à chaque cluster [3]. Le problème du clustering peut être défini comme un problème d'optimisation mono-objectif avec une seule fonction objective (OF) présentée par un critère de validité de cluster CVI. En tenant compte du fait que les problèmes de clustering sont devenus plus complexes et que différents CVI sont apparus, il est donc nécessaire de considérer le clustering comme un problème d'optimisation multi-objectif (MOCO) avec différentes OF présentées par des CVI.

Dans la littérature, diverses techniques de MOCO ont été proposées, notamment les algorithmes MOCK, VGAPS, GenClustMOO, GenClustPESA2 et cOptBees-MO.

Dans cette étude un nouvel algorithme a été proposé pour résoudre le problème du MOCO, nommé MO-β-SSA, cet algorithme est la version multi-objective de l'algorithme β-SSA proposé par Rokbani et al. [5].

Dans la suite de ce papier, on rappelle l'algorithme β-SSA, puis on présente brièvement sa variante multi-objectif appliquée au clustering. Les résultats préliminaires figurent au paragraphe 3, suivie de la conclusion et des perspectives.

2 Le β-salpedia pour le clustering

2.1 L'algorithme d'essaim de slapidia β

L'algorithme β-SSA [5] est une variation de SSA qui implique l'utilisation de la fonction bêta pour distribuer les salpedia leaders. Cette fonction peut être modulée pour contrôler la distribution des leaders dans tout l'espace de recherche, elle offre de multiples options pour les gérer dans le cadre d'une stratégie d'exploration/exploitation.

Le contrôle du comportement s'apparente à une modification de l'outil de recherche utilisant la capacité de la fonction bêta à approximer plusieurs types de distributions. Deux profils de distributions sont utilisés : le premier profil est un ***profile d'exploration β1*** ; le deuxième est un ***profile d'exploitation β2***. A l'exécution de β-SSA, un test itératif évalue chaque leader, si Fitness (S_i) est >= min-fitness donc la mise à jour de la position de S_i se fait à travers ***profile d'exploration β1*** en utilisant l'Eq.(1), sinon la mise à jour se fait à travers ***profile d'exploitation β2*** en utilisant l'Eq.(2). Si la valeur admissible min_fitness est inconnue la moyenne des fitness des leaders est utilisée.

Tableau 1: Comparaison par la métrique F-measure (F-M) sur des ensembles de données de référence UCI, les meilleures valeurs de F-Measure sont marquées en gras

Dataset	NC	GenClust MOO		GenClust PESA2		MOCK		VGAPS		cOptBee-MO		K-means		Spectral-Clustering		MO-β-SSA	
		K	F-M	K	F-M	K	F-M	K	F-M	K	F-M	K	F-M	K	F-M	K	F-M
Iris	3	3	0.79	3	0.93	2	0.78	3	0.76	3	**1.00**	3	0.88	3	0.90	3	**1.00**
Cancer	2	2	0.97	2	**0.98**	2	0.82	2	0.95	3	0.94	2	0.92	2	0.65	2	0.90
Newthyroid	3	3	0.86	9	0.69	2	0.74	5	0.66	4	0.86	3	0.86	3	0.66	3	**0.88**
LiverDisorder	2	2	0.67	5	0.60	2	0.67	2	0.70	2	0.67	2	0.64	2	0.67	2	**0.72**
Glass	6	6	0.49	5	0.53	5	0.53	5	0.53	2	**0.88**	6	0.56	6	0.53	6	0.59
Wine	3	3	0.71	1	0.44	3	0.73	6	0.62	2	0.65	3	0.71	3	0.50	3	**0.85**

Tableau 2: Comparaison par la métrique NMI sur des ensembles de données de référence UCI, les meilleures valeurs de NMI sont marquées en gras.

Dataset	K-means	Spectral-Clustering	MO-β-SSA
	NMI	NMI	NMI
Iris	0.74	0.80	**1.00**
Cancer	**0.63**	0.06	0.50
Newthyroid	0.49	0.13	**0.63**
Liver Disorder	0.0011	0.008	**0.14**
Glass	0.41	0.35	**0.47**
Wine	0.42	0.03	**0.76**

Tableau 3: Comparaison par la métrique ARI sur des ensembles de données de référence UCI, les meilleures valeurs de ARI sont marquées en gras.

Dataset	K-means	Spectral-Clustering	MO-β-SSA
	ARI	ARI	ARI
Iris	0.71	0.75	**1.00**
Cancer	**0.72**	-0.0014	0.61
Newthyroid	0.57	0.10	**0.66**
Liver Disorder	-0.0063	-0.0032	**0.20**
Glass	0.27	0.18	**0.28**
Wine	0.37	-0.003	**0.68**

Si (Fitness (Si) > = min-fitness)

//utiliser profile d'exploration β1

$$X_j^1 = \begin{cases} Fj + (rand\_\beta_1^{-1}(ubj - lbj) + lbj) \; if \; c3 \geq 0.5 \\ Fj - \left(rand_{\beta_1}^{-1}(ubj - lbj) + lbj\right) \; if \; c3 < 0.5 \end{cases} \quad (1)$$

Sinon

//utiliser profile d'exploitation β2

$$X_j^1 = \begin{cases} Fj + (rand\_\beta_2^{-1}(ubj - lbj) + lbj) \; if \; c3 \geq 0.5 \\ Fj - \left(rand_{\beta_2}^{-1}(ubj - lbj) + lbj\right) \; if \; c3 < 0.5 \end{cases} \quad (2)$$

Fin_Si

La formulation de bêta est comme suit:

$$\beta(x; p, q, x_0, x_1) = \begin{cases} \left(\frac{x - x_0}{x_c - x_0}\right)^p \left(\frac{x_1 - x_0}{x_1 - x_1}\right)^q & if \; x \in]x_0, x_1[\\ 0 & elswhere \end{cases} \quad (3)$$

Avec $p, q, (x_0 < x_1) \in \mathbb{R}$, et

$$x_c = \frac{p.x_1 + q.x_0}{p+q} \quad (4)$$

2.2 Le multi-objectif β salpedia pour le clustering (MO-β-SSA)

MO-β-SSA est la forme multi-objectif de l'algorithme β-SSA, ou la dominance est utilisée pour sélectionner une solution déposée dans un référentiel dit "solutions de l'ensemble de **pareto**"[2]. La solution étant le candidat le plus proche du Point-Utopie [1].

Dans MO-β-SSA pour clustering, un salpedia S_i est une matrice (k*d); k et d présentent respectivement le nombre de clusters et le nombre de caractéristiques liées à la BD; chaque vecteur-ligne de cette matrice présente un centre de cluster. Les entrées de l'algorithme sont : la BD, les paramètres relatifs à MO-β-SSA et les OF ; en sorties l'algorithme produit une solution de clustering représentée par : le nombre de clusters détectés, les centres des clusters et les données groupées.

3 Résultats

Pour la validation expérimentale, MO-β-SSA a été appliqué sur 6 BD: Iris [4], Cancer [4], Newthyroid [4], Wine [4], LiverDisorder [4] et Glass [4] ; il a été exécuté 30 fois avec les paramètres d'entrée suivants : Nombre de salpedia =50, Taille du référentiel = 25 et Nombre d'itérations = 100. Les métriques F-Mesure [1][2][6], Normalised Mutual Information (NMI) [6] et Adjusted Rand Index (ARI) [6] sont utilisée pour évaluer et comparer les solutions finales de clustering obtenues par MO-β-SSA et ses concurrents.

Tableaux 1,2 et 3 illustrent les résultats MO-β-SSA et ses concurrents. Dans le tableau 1 , NC désigne le nombre réel de clusters et k représente le nombre de clusters détecté par chaque algorithme. Les meilleures valeurs de F-Measure, NMI et ARI sont marquées en gras. MO-β-SSA a eu une performance compétitive par rapport à ses concurrents, il a également détecté le nombre correct de clusters pour toutes les BD.

4 Conclusions et perspectives

Dans cet article, **MO-β-SSA** a été proposé pour étudier l'optimisation multi-objectif du clustering. L'évaluation des performances de **MO-β-SSA** est basée sur l'utilisation

de 3 CVI (I-index [1], Con-index [1] et Sym-index [1]) comme OF, 6 BD pour le test et 7 algorithmes pour la comparaison: 5 algorithmes de clustering multi- objectifs (GenClustMOO, MOCK, VGAPS, GenClustPESA2et cOptBees-MO) et deux algorithmes standard de clustering mono-objectif (Kmeans et Clustering Spectral), en utilisant les métriques F-mesure, NMI, ARI. **MO-β-SSA** s'avère intéressant sur cette première investigation limitée. Une étude étendue sur des bases plus conséquentes et une comparaison par rapport à d'autres algorithmes standard de clustering sont nécessaires (i.e. DEC, iDEC,...) sont nécessaires. Nous envisageons également une étude approfondie du profilage bêta des phases de recherche et d'exploitation.

Bibliographie

1. Cunha, D., Cruz, D., Politi, A., de Castro, L. N., & Maia, R. D. (2017). Bio-inspired multiobjective clustering optimization: A survey and a proposal. Artificial Intelligence Research, 6(2), 10-24.
2. Deb, K. (2011). Multi-objective optimisation using evolutionary algorithms: an introduction (pp. 3-34). Springer London.
3. Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8), 651-666.
4. Lichman, M. (2013). UCI machine learning repository, 2013.
5. Rokbani, N., Mirjalili, S., Slim, M., & Alimi, A. M. (2022). A beta salp swarm algorithm meta-heuristic for inverse kinematics and optimization. *Applied Intelligence*, *52*(9), 10493-10518.
6. Rezaei, M., & Fränti, P. (2016). Set matching measures for external cluster validity. *IEEE transactions on knowledge and data engineering*, *28*(8), 2173-2186.