

# Improving Group Search Optimization Through Local Search Heuristics for Automatic Data Clustering

Luciano D. S. Pacífico<sup>1</sup>, Teresa B. Ludermir<sup>2</sup>

<sup>1</sup>Departamento de Computação – Universidade Federal Rural de Pernambuco  
R. Dom Manuel de Medeiros, S/N – 52.171-900 – Recife – PE – Brazil

<sup>2</sup>Centro de Informática – Universidade Federal de Pernambuco  
Av. Jornalista Aníbal Fernandes, – 50740-560 – Recife – PE – Brazil

luciano.pacifico@ufrpe.br, tbl@cin.ufpe.br

**Abstract.** *In this work, three Automatic Clustering models, based on Group Search Optimization (GSO) meta-heuristic, are introduced, named RHGSO, ADHGSO and BDHGSO. In proposed models, GSO global search capability is enhanced by means of local search heuristic operators adapted to the context of Automatic Data Clustering, where cluster activation, deactivation and centroid replacement are executed, seeking out to promote perturbations that would speedup the exploration performed by GSO group. The proposed GSO-based algorithms are compared to other Evolutionary and Swarm Intelligence partitioned Automatic Clustering models from the literature, showing promising results.*

**Resumo.** *Neste trabalho, três modelos de Agrupamento Automático de Dados, baseados na meta-heurística de Otimização por Busca em Grupo (GSO), são introduzidos, chamados RHGSO, ADHGSO e BDHGSO. Nos modelos propostos, a busca global do GSO é melhorada através de heurísticas de busca local adaptadas ao contexto de Agrupamento Automático de Dados, onde operações de ativação, desativação e substituição de centroides de agrupamentos são executadas, objetivando a realização de perturbações que visam o aumento da velocidade de exploração do grupo do GSO. Os algoritmos propostos são comparados a outros Algoritmos Evolucionários e de Inteligência de Enxames da literatura, apresentando resultados promissores.*

## 1. Introdução

A Análise de Agrupamentos tem se destacado como uma poderosa ferramenta para a análise exploratória e categorização de dados não rotulados, sendo uma das principais atividades de Aprendizagem de Máquina Não Supervisionada. A Análise de Agrupamentos vem sendo utilizada em diversos campos de pesquisa, como Bioinformática, Ciências Sociais, Sistemas de Recomendação, Análise de Imagens, dentre outras aplicações, uma vez que há uma demanda cada vez maior por ferramentas computacionais que sejam capazes de processar enormes quantidades de dados de forma rápida e precisa, e que auxiliem pesquisadores e empresas na descoberta de padrões e tendências existentes em suas bases de dados.

Os algoritmos de agrupamento buscam organizar conjuntos de amostras (bases de dados) em grupos (ou agrupamentos) naturais, de acordo com relações de proximidade

e semelhança inerentes a esses dados, sendo essa tarefa de separação e organização das amostras considerada uma tarefa complexa de Otimização Combinatória (problema NP-Difícil). Os algoritmos mais populares de Agrupamento de Dados se encontram na categoria de algoritmos de agrupamento particionais, sendo caracterizados pela realização de uma divisão iterativa dos dados em um número pré-definido de agrupamentos (parâmetro dos algoritmos particionais), no intuito da geração de uma partição final que otimize uma função objetivo. Porém, os algoritmos particionais tradicionais possuem várias limitações, como a inexistência de mecanismos de busca global, sua sensibilidade à inicialização, além da necessidade de conhecimento prévio sobre o problema a ser resolvido para a correta determinação do número final de agrupamentos desejado.

Nas últimas décadas, meta-heurísticas inspiradas na natureza, tais quais os Algoritmos Evolucionários (EAs), como os Algoritmos Genéticos [Holland 1992], Evolução Diferencial [Storn and Price 1995] e a Otimização por Busca com *Backtracking* [Civicioglu 2013], e os algoritmos de Inteligência de Enxames (SIs), como a Otimização por Enxame de Partículas [Kennedy and Eberhart 1995], a Otimização por Colônia de Formigas [Dorigo et al. 1996], e a Otimização por Busca em Grupo [He et al. 2009], têm sido frequentemente aplicadas à realização da tarefa de Agrupamento de Dados, uma vez que esses modelos são conhecidos por suas boas capacidades de realização de busca globais, assim como por seus mecanismos de fuga de pontos ótimos locais. Nos EAs e SIs, um conjunto de soluções (população) é mantido e melhorado ao longo de um processo geracional (iterativo), no qual operadores evolucionários são executados a essas soluções em uma tentativa de melhorar o valor de uma função objetivo (função de *fitness*). Os EAs e SIs adaptados ao contexto de Análise de Agrupamentos oferecem boas respostas ao problema de sensibilidade à inicialização e da falta de bons mecanismos de busca global de algoritmos particionais tradicionais, sendo esses métodos comumente usados também para a estimação do número ideal de agrupamentos da partição final dos dados, em uma tarefa conhecida por Agrupamento Automático de Dados [Das et al. 2007, José-García and Gómez-Flores 2016, Pacifico and Ludermir 2020, Ezugwu et al. 2021].

Em decorrência da natureza estocástica dos EAs e SIs, esses modelos tendem a ser bastante lentos para convergir, o que tem ocasionado a surgimento de vários modelos híbridos entre EAs (e SIs) e heurísticas voltadas ao contexto das aplicações [Ikotun et al. 2021, Pacifico and Ludermir 2021]. Visando o aumento da velocidade de exploração do espaço de busca realizada por meta-heurísticas inspiradas na natureza, este trabalho propõe três novos modelos baseados na Otimização por Busca em Grupo (GSO), chamados RHGSO, ADHGSO e BDHGSO, que fazem uso de diferentes heurísticas de modificação da população do GSO através de perturbações, representadas por operações de ativação, desativação e substituição de centroides de agrupamentos, para a realização da tarefa de Agrupamento Automático de Dados. Os modelos propostos buscam a otimização simultânea do número final de grupos estimados, assim como a melhor distribuição dos padrões nesses grupos. O GSO é escolhido por ter apresentado boas performances quando comparado a outros algoritmos de Computação Evolucionária, como o PSO e o GA [He et al. 2009].

O trabalho está dividido como segue. A Seção 2 apresenta o GSO. Após isso, os modelos propostos são descritos (Seção 3). A avaliação experimental é realizada na Seção 4, seguida pelas conclusões e tendências para pesquisas futuras (Seção 5).

## 2. Otimização por Busca em Grupo (GSO)

O GSO é um SI baseado em uma tentativa de simulação do comportamento de busca de animais sociais, e na teoria dos grupos vivos (*living group theory*). No GSO, a população  $G$  de  $S$  indivíduos é chamada de *grupo*, enquanto os indivíduos da população são chamados de *membros*. A busca desempenhada por cada membro no GSO leva em consideração o campo de varredura visual do mesmo. Em um problema definido por um espaço de busca  $n$ -dimensional, o  $i$ -ésimo membro na  $t$ -ésima iteração da busca terá uma posição atual definida pelo vetor  $\mathbf{X}_i^t \in \mathfrak{R}^n$  e um ângulo de cabeça definido pelo vetor  $\alpha_i^t \in \mathfrak{R}^{n-1}$ . A direção da busca do  $i$ -ésimo membro, que é um vetor  $\mathbf{D}_i^t(\alpha_i^t) = (d_{i1}^t, \dots, d_{in}^t)$ , pode ser calculada por  $\alpha_i^t$  através de uma transformação polar para coordenadas cartesianas dada por:

$$\begin{aligned} d_{i1}^t &= \prod_{q=1}^{n-1} \cos(\alpha_{iq}^t), \\ d_{ij}^t &= \sin(\alpha_{i(j-1)}^t) \prod_{q=1}^{n-1} \cos(\alpha_{iq}^t) (j=1, \dots, n-1), \\ d_{in}^t &= \sin(\alpha_{i(n-1)}^t) \end{aligned} \quad (1)$$

Um grupo no GSO consiste de três tipos de membros: *producers*, *scroungers* e *rangers*, sendo os *rangers* uma modificação ao modelo PS original proposta pelo GSO. Durante cada geração do GSO, o membro do grupo que encontrou a melhor reserva de recursos (melhor valor de *fitness*), é escolhido como *producer*. O *producer* executará uma estratégia de varredura baseada em seu campo de visão. Essa estratégia recebe o nome de *producing*. Na  $t$ -ésima geração de uma execução do GSO, o *producer*  $\mathbf{X}_p^t$  varrerá o espaço de busca do problema através da observação de três pontos aleatórios em seu campo visual: um ponto a zero grau ( $\mathbf{X}_z$ ), um ponto no hipercubo à sua direita ( $\mathbf{X}_r$ ), e um ponto no hipercubo à sua esquerda ( $\mathbf{X}_l$ ), de acordo com eq. (2).

$$\mathbf{X}_z = \mathbf{X}_p^t + r_1 l_{max} \mathbf{D}_p^t(\alpha_p^t), \quad \mathbf{X}_r = \mathbf{X}_p^t + r_1 l_{max} \mathbf{D}_p^t(\alpha_p^t + \frac{\mathbf{r}_2 \theta_{max}}{2}), \quad \mathbf{X}_l = \mathbf{X}_p^t + r_1 l_{max} \mathbf{D}_p^t(\alpha_p^t - \frac{\mathbf{r}_2 \theta_{max}}{2}) \quad (2)$$

onde  $r_1 \in \mathfrak{R}$  é um número aleatório obtido através de uma distribuição normal com média 0 (zero) e desvio padrão 1 (um),  $\mathbf{r}_2 \in \mathfrak{R}^{n-1}$  é uma sequência uniforme obtida aleatoriamente no intervalo (0, 1),  $\theta_{max} \in \mathfrak{R}^{n-1}$  é o ângulo máximo de busca e  $l_{max} \in \mathfrak{R}$  é a distância máxima de busca, definida pela equação eq. (3):

$$l_{max} = \|\mathbf{U} - \mathbf{L}\| = \sqrt{\sum_{k=1}^n (U_k - L_k)^2} \quad (3)$$

onde  $U_k$  e  $L_k$  denotam o limite superior e o limite inferior da  $k$ -ésima dimensão do problema, respectivamente.

Todos os *scroungers* tentarão alcançar as reservas já encontradas pelo *producer*, sendo essa estratégia denominada *scrounging*. O operador de *scrounging* no GSO é dado por (eq. (4)):

$$\mathbf{X}_i^{t+1} = \mathbf{X}_i^t + \mathbf{r}_3 \circ (\mathbf{X}_p^t - \mathbf{X}_i^t) \quad (4)$$

onde  $\mathbf{r}_3 \in \mathfrak{R}^n$  é uma sequência uniforme aleatória obtida no intervalo (0, 1), e  $\circ$  é o produto de Hadamard ou Schur, que calcula o produto interno entre dois vetores.

Os *rangers* irão executar buscas aleatórias no espaço do problema, sendo essa estratégia denominada *ranging*. O operador de *ranging* é apresentado abaixo (eq. (5)).

$$\mathbf{X}_i^{t+1} = \mathbf{X}_i^t + l_i \mathbf{D}_i^t(\alpha_i^{t+1}), \quad l_i = ar_1 l_{max} \quad (5)$$

O GSO define uma estratégia de contenção aos membros que, em decorrência da atualização de suas posições pelos operadores evolucionários da estratégia, acabam sendo posicionados fora dos limites do espaço de busca do problema. No GSO tradicional, quando um membro escapa dos limites do espaço de busca após a atualização de sua posição, o mesmo será reconduzido à sua posição anterior dentro dos limites do espaço de busca do problema. A importância do tratamento de indivíduos da população que extrapolam os limites do espaço de busca do problema está relacionada ao fato de que soluções inválidas (que não se enquadrariam às restrições do problema tratado) poderiam ser geradas e consideradas como soluções possíveis para o problema. O GSO é apresentado no Algoritmo 1.

---

### Algorithm 1 GSO

---

```

 $t \leftarrow 0.$ 
Inicialize aleatoriamente as posições dos membros  $\mathbf{X}_i^{(0)} \in G$ . Determine aleatoriamente os ângulos de cabeça  $\alpha_i^{(0)}$  de cada membro  $\mathbf{X}_i^{(0)} \in G^{(0)}$ 
Calcule  $f(\mathbf{X}_i^{(0)})$  para cada membro  $\mathbf{X}_i^{(0)}$ .
enquanto (as condições de término não forem satisfeitas) faça
  Use o melhor membro do grupo ( $\mathbf{X}_p^t$ ) para a execução do producing (eq. (2)).
  Escolha um percentual dos membros restantes para a execução da estratégia de scrounging.
  Ranging: Os membros restantes executarão ranging por buscas aleatórias no espaço do problema.
  Aplique o operador de controle do GSO para os membros que escapam das dimensões do espaço de busca do problema.
  Calcule o novo fitness  $f(\mathbf{X}_i^{t+1})$  para cada membro  $\mathbf{X}_i^{t+1}$ .
   $t \leftarrow t + 1.$ 
fim enquanto
retorne  $\mathbf{X}_p^{t_{max}}$ .

```

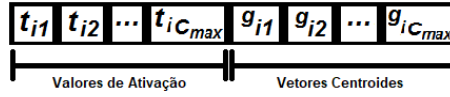
---

### 3. Abordagens Propostas: RHGSO, ADHGSO e BDHGSO

Nesta seção, as abordagens de agrupamento automático propostas (ou seja, o RHGSO, o ADHGSO e o BDHGSO) são descritas. Para as três variações do GSO propostas, a busca global realizada pelo GSO é melhorada através de operadores heurístico de busca local, visando a realização de perturbações nos membros do grupo do GSO, a cada geração, semelhantes aos operadores de mutação de algoritmos como os Algoritmos Genéticos, porém tais perturbações são ajustadas ao contexto de Agrupamento Automático de Dados, promovendo, desta forma, uma maior variabilidade e diversidade aos membros do grupo do GSO.

A modelagem e representação do problema de Agrupamento Automático de Dados adotada é semelhante à forma mais comumente adotada na literatura [Das et al. 2007, José-García and Gómez-Flores 2016, Pacifico and Ludermir 2020]: cada membro do GSO representará um conjunto de centroides de agrupamento, assim como valores de ativação que indicarão se um determinado centroide será levado em consideração ou não na definição da partição representada pelo membro do grupo em análise. Formalmente, considere uma partição  $P_C$  da base de dados, que contém  $N$  padrões  $\mathbf{o}_j \in \mathbb{R}^m$  ( $j = 1, 2, \dots, N$ ), em no máximo  $C_{max}$  agrupamentos. Cada agrupamento será representado por seu centroide  $\mathbf{g}_c \in \mathbb{R}^m$  ( $c = 1, 2, \dots, C_{max}$ ). Cada membro  $\mathbf{X}_i \in \mathbb{R}^n$  (onde  $n = C_{max} + C_{max} \times m$ ) no grupo  $G$  representa  $C_{max}$  valores de ativação e  $C_{max}$  centroides de agrupamentos ao mesmo tempo (vide Fig. 1).

Na  $t$ -ésima geração, o membro  $\mathbf{X}_i^t$  terá seu valor de *fitness* avaliado considerando apenas os agrupamentos **ativados**, ou seja, os agrupamentos cujos valores de



**Figura 1. Codificação dos membros do GSO: os primeiros  $C_{max}$  genes representam os valores de ativação para cada agrupamento candidato, enquanto os  $C_{max} \times m$  genes seguintes representam os  $C_{max}$   $m$ -dimensionais centroides dos agrupamentos candidatos.**

ativação sejam tais que  $t_{ic}^t \geq 0.5$ . A função de *fitness* escolhida deve ser apropriada para a realização da tarefa de Agrupamentos Automático de Dados, de modo a otimizar, simultaneamente, o número de agrupamentos finais, e também os vetores centroides [José-García and Gómez-Flores 2016].

A inicialização é realizada através da escolha aleatória de  $C_{max}$  padrões da base de dados em análise para compor os centroides iniciais de cada um dos membros  $\mathbf{X}_i^{(0)}$ , e, de modo semelhante, os valores de ativação  $t_{ic}^{(0)}$  (onde  $c = 1, 2, \dots, C_{max}$ ) são obtidos aleatoriamente a partir de uma distribuição uniforme no intervalo  $U(0, 1)$ . Após a inicialização e a avaliação de cada membro em  $G^{(0)}$ , o processo geracional terá início, para cada uma das abordagens propostas, de modo semelhante ao GSO tradicional. Na  $t$ -ésima geração, o membro mais promissor  $\mathbf{X}_p^t$  é escolhido para a execução do operador de *producing* (eq. (2)). Em seguida, um percentual dos  $\mathbf{X}_i^t$  membros restantes são escolhidos para executarem a estratégia de *scrounging* (eq. (4)). Por fim, os membros restantes serão escolhidos como *rangers*, executando o *ranging* (eq. (5)). O operador de correção de posições que extrapolaram as dimensões do espaço de busca do problema é executado em seguida, de modo semelhante ao GSO. Após a execução dos operadores evolucionários do GSO, na

---

## Algorithm 2 Heurística RHGSO

---

Escolha, para o membro  $\mathbf{X}_i^{t+1}$  atual, com a mesma probabilidade entre cada um, executar um dos operadores abaixo:

**Ativação de um Centroide Aleatório:**

Selecione um valor de ativação aleatório  $t_{ic}^{t+1}$ , tal que  $t_{ic}^{t+1} < 0.5$  (indicação de que o centroide  $\mathbf{g}_{ic}^{t+1}$  está **inativo**).

Faça  $t_{ic}^{t+1} = t_{ic}^{t+1} + 0.5$ .

**Desativação de um Centroide Aleatório:**

Selecione um valor de ativação aleatório  $t_{ic}^{t+1}$ , tal que  $t_{ic}^{t+1} \geq 0.5$  (indicação de que o centroide  $\mathbf{g}_{ic}^{t+1}$  está **ativo**).

Faça  $t_{ic}^{t+1} = t_{ic}^{t+1} - 0.5$ .

**Substituição de um Centroide Aleatório:**

Selecione um valor de ativação aleatório  $t_{ic}^{t+1}$ , tal que  $t_{ic}^{t+1} > 0.5$  (indicação de que o centroide  $\mathbf{g}_{ic}^{t+1}$  está **ativo**).

Substitua o centroide  $\mathbf{g}_{ic}^{t+1}$  por um padrão aleatório da base de dados (reinicialização aleatória).

Atualize  $\mathbf{X}_i^{t+1}$  com os novos valores de ativação e centroides de agrupamento.

---

$t$ -ésima geração e com uma probabilidade  $m_r^t$ , cada membro  $\mathbf{X}_i^t$  do grupo  $G^t$  executará, aleatoriamente, um dos seguintes operadores de perturbação: **ativação de um centroide atualmente inativo**, **desativação de um centroide atualmente ativo**, e na **substituição aleatória de um centroide atualmente ativo**. A forma como a heurística de perturbação local é implementada varia de acordo com cada modelo, sendo descritas no Algoritmo 2 (RHGSO), Algoritmo 3 (ADHGSO), e Algoritmo 4 (BDHGSO). Nos modelos propostos, a probabilidade de execução das perturbações decai linearmente com o número de gerações, uma vez que, no início da busca executada por cada algoritmo, é desejável que os mesmos realizem explorações mais amplas do espaço de busca do problema, tentando identificar regiões mais promissoras nesse espaço; porém, também é desejável que nas últimas gerações, os membros do grupo foquem seus esforços na exploração das regiões

promissoras do espaço de busca identificadas até o momento, sem grandes perturbações aleatórias, gerando mais estabilidade ao algoritmo.

---

### Algorithm 3 Heurística ADHGSO

---

Escolha, para o membro  $\mathbf{X}_i^{t+1}$  atual, com a mesma probabilidade entre cada um, executar um dos operadores abaixo:

**Ativação do Centroide com Maior Distância Média:**

Encontre o valor de ativação  $t_{ic}^{t+1}$  **inativo** (ou seja,  $t_{ic}^{t+1} < 0.5$ ), correspondente ao centroide  $\mathbf{g}_{ic}^{t+1}$  cuja **distância média** em relação aos centroides ativos  $\mathbf{g}_{ij}^{t+1}$  de  $\mathbf{X}_i^{t+1}$  (ou seja, aos centroides cujo valor de ativação seja tal que  $t_{ij}^{t+1} \geq 0.5$ ) seja **a maior**.

Faça  $t_{ic}^{t+1} = t_{ic}^{t+1} + 0.5$ .

**Desativação do Centroide com Menor Distância Média:**

Encontre o valor de ativação  $t_{ic}^{t+1}$  **ativo** (ou seja,  $t_{ic}^{t+1} \geq 0.5$ ), correspondente ao centroide  $\mathbf{g}_{ic}^{t+1}$  cuja **distância média** em relação aos demais centroides ativos  $\mathbf{g}_{ij}^{t+1}$  de  $\mathbf{X}_i^{t+1}$  (ou seja, aos centroides cujo valor de ativação seja tal que  $t_{ij}^{t+1} \geq 0.5$ ) seja **a menor**.

Faça  $t_{ic}^{t+1} = t_{ic}^{t+1} - 0.5$ .

**Substituição do Centroide com Menor Distância Média:**

Encontre o valor de ativação  $t_{ic}^{t+1}$  **ativo** (ou seja,  $t_{ic}^{t+1} \geq 0.5$ ), correspondente ao centroide  $\mathbf{g}_{ic}^{t+1}$  cuja **distância média** em relação aos demais centroides ativos  $\mathbf{g}_{ij}^{t+1}$  de  $\mathbf{X}_i^{t+1}$  (ou seja, aos centroides cujo valor de ativação seja tal que  $t_{ij}^{t+1} \geq 0.5$ ) seja **a menor**.

Substitua o centroide  $\mathbf{g}_{ic}^{t+1}$  por um padrão aleatório da base de dados (reinicialização aleatória).

Atualize  $\mathbf{X}_i^{t+1}$  com os novos valores de ativação e centroides de agrupamento.

---



---

### Algorithm 4 Heurística BDHGSO

---

Escolha, para o membro  $\mathbf{X}_i^{t+1}$  atual, com a mesma probabilidade entre cada um, executar um dos operadores abaixo:

**Ativação do Centroide com Maior Distância Mínima:**

Encontre o valor de ativação  $t_{ic}^{t+1}$  **inativo** (ou seja,  $t_{ic}^{t+1} < 0.5$ ), correspondente ao centroide  $\mathbf{g}_{ic}^{t+1}$  cuja **menor distância** em relação aos centroides ativos  $\mathbf{g}_{ij}^{t+1}$  de  $\mathbf{X}_i^{t+1}$  (ou seja, aos centroides cujo valor de ativação seja tal que  $t_{ij}^{t+1} \geq 0.5$ ) seja **a maior**.

Faça  $t_{ic}^{t+1} = t_{ic}^{t+1} + 0.5$ .

**Desativação do Centroide com Menor Distância Máxima:**

Encontre o valor de ativação  $t_{ic}^{t+1}$  **ativo** (ou seja,  $t_{ic}^{t+1} \geq 0.5$ ), correspondente ao centroide  $\mathbf{g}_{ic}^{t+1}$  cuja **maior distância** em relação aos demais centroides ativos  $\mathbf{g}_{ij}^{t+1}$  de  $\mathbf{X}_i^{t+1}$  (ou seja, aos centroides cujo valor de ativação seja tal que  $t_{ij}^{t+1} \geq 0.5$ ) seja **a menor**.

Faça  $t_{ic}^{t+1} = t_{ic}^{t+1} - 0.5$ .

**Substituição do Centroide com Menor Distância Máxima:**

Encontre o valor de ativação  $t_{ic}^{t+1}$  **ativo** (ou seja,  $t_{ic}^{t+1} \geq 0.5$ ), correspondente ao centroide  $\mathbf{g}_{ic}^{t+1}$  cuja **maior distância** em relação aos demais centroides ativos  $\mathbf{g}_{ij}^{t+1}$  de  $\mathbf{X}_i^{t+1}$  (ou seja, aos centroides cujo valor de ativação seja tal que  $t_{ij}^{t+1} \geq 0.5$ ) seja **a menor**.

Substitua o centroide  $\mathbf{g}_{ic}^{t+1}$  por um padrão aleatório da base de dados (reinicialização aleatória).

Atualize  $\mathbf{X}_i^{t+1}$  com os novos valores de ativação e centroides de agrupamento.

---

Nos modelos propostos, se a qualquer momento da busca um membro  $\mathbf{X}_i$  representar menos de dois agrupamentos ativos, o mesmo será reinicializado aleatoriamente, através de um processo semelhante ao adotado na etapa de inicialização do grupo. As abordagens propostas são sintetizadas no Algoritmo 5.

## 4. Análise Experimental

Nesta seção, os modelos propostos (RHGSO, ADHGSO e BDHGSO) são testados e comparados a seis outros EAs e SIs da literatura, através da avaliação de nove bases de dados reais, obtidas através do *UCI Machine Learning Repository* [Asuncion and Newman 2007]. As bases de dados selecionadas são apresentadas na Tabela 1.

Os EAs e SIs selecionados para a análise comparativa são: O Algoritmo Genético (GA), a Evolução Diferencial (DE), a Otimização por Enxame de Partículas (PSO), a Otimização por Busca com *Backtracking* (BSA), o GSO tradicional e um modelo híbrido

---

**Algorithm 5** Modelos Propostos (RHGSO, ADHGSO e BDHGSO)
 

---

$t \leftarrow 0$ .

**Inicialização:** para cada membro  $\mathbf{X}_i^{(0)} \in G^{(0)}$ , escolha aleatoriamente  $C_{max}$  padrões da base de dados como os centroides de agrupamento iniciais  $\mathbf{g}_{ic}$  ( $c = 1, 2, \dots, C_{max}$ ). Determine aleatoriamente os valores de ativação  $t_{ic}^{(0)}$  e os ângulos de cabeça  $\alpha_i^{(0)}$  de cada membro  $\mathbf{X}_i^{(0)} \in G^{(0)}$ . Para cada membro  $\mathbf{X}_i^{(0)}$ , encontre o agrupamento correspondente à cada padrão  $\mathbf{o}_j$  pelo critério de menor distância.

**Calcule** o valor do *fitness* ( $f(\mathbf{X}_i^{(0)})$ ) para cada membro  $\mathbf{X}_i^{(0)}$ .

**enquanto** (as condições de término não forem satisfeitas) **faça**

$m_r^t \leftarrow m_r^{(0)}(t_{max} - t)/(t_{max})$ .

**Use** o melhor membro do grupo ( $\mathbf{X}_p^t$ ) para a execução do *producing* (eq. (2)). Para cada ponto gerado, determine a partição formada pelo mesmo pela atribuição de cada padrão da base de dados ao agrupamento mais próximo.

**Escolha** um percentual dos membros restantes para a execução da estratégia de *scrounging*.

**Rangig:** Os membros restantes executarão *rangig* por buscas aleatórias no espaço do problema.

**Aplique** o operador de controle do GSO para os membros que escapam das dimensões do espaço de busca do problema.

**Execute** uma única etapa do K-Means para refinar cada membro do novo grupo  $G^{t+1}$ .

**Reinicialize** aleatoriamente cada membro no novo grupo  $G^{t+1}$  que contenha menos que dois agrupamentos ativos.

**Calcule** o valor do *fitness* ( $f(\mathbf{X}_i^{t+1})$ ) de cada membro no novo grupo  $G^{t+1}$ .

**para** (cada membro  $\mathbf{X}_i^{t+1} \in G^{t+1}$ ) **faça**

**se** (RHGSO) **então**

**Execute**, com probabilidade  $m_r$ , a heurística descrita no Algoritmo 2) para  $\mathbf{X}_i^{t+1}$ .

**senão se** (ADHGSO) **então**

**Execute**, com probabilidade  $m_r$ , a heurística descrita no Algoritmo 3) para  $\mathbf{X}_i^{t+1}$ .

**senão** // BDHGSO

**Execute**, com probabilidade  $m_r$ , a heurística descrita no Algoritmo 4) para  $\mathbf{X}_i^{t+1}$ .

**fim.se**

**fim.para**

$t \leftarrow t + 1$ .

**fim.enquanto**

**Retorne**  $\mathbf{X}_p^{t_{max}}$ .

---

**Tabela 1. Bases de Dados Reais.**

Base de Dados	Padrões	Características	Classes
<i>Banknote Authentication</i>	1372	4	2
<i>Cancer</i>	699	9	2
<i>Diabetes</i>	768	8	2
<i>Heart</i>	270	13	2
<i>Ionosphere</i>	351	34	2
<i>Iris</i>	150	4	3
<i>Page Blocks Classification</i>	5473	10	5
<i>Seeds</i>	210	7	3
<i>Waveform</i>	5000	21	3

**Tabela 2. Hiperparâmetros para os Algoritmos.**

Algoritmo	Hiperparâmetro	Valor
Todos os Algoritmos	$t_{max}$	200
	$S$	100
	$C_{max}$	20
GA	Taxa de Cruzamento	0.8
	Taxa de Mutação	0.1
	Taxa de Seleção	0.8
DE	$F$	0.8
	Taxa de Cruzamento	0.9
PSO	$c_1$	2.0
	$c_2$	2.0
	$w$	0.9 até 0.4
GSO, BGSO, RHGSO, ADHGSO e BDHGSO	Taxa de <i>Scroungers</i>	0.8
	$\theta_{max}$	$\pi/a^2$
	$\alpha_0$	$\pi/4$
	$\alpha_{max}$	$\theta_{max}/2$
BSA e BGSO	$mixrate$	1
	$F$	$3N(0,1)$
RHGSO, ADHGSO e BDHGSO	$m_r$	0.9

entre GSO e BSA [Pacífico and Ludermitr 2020]. Tais algoritmos representam algumas das melhores abordagens do estado da arte em Computação Evolucionária e Inteligência de Enxames, tendo sido aplicados na solução de vários problemas práticos [Preetha 2021, Shi et al. 2020, Ye and Zheng 2021, Pacífico and Ludermitr 2020, Pacífico 2020]. Todos os EAs e SIs adotados para comparações foram adaptados ao contexto de Agrupamento Automático de Dados, de acordo com uma modelagem semelhante à adotada para os modelos propostos (vide Seção 3). Os hiperparâmetros de configuração dos modelos testados são apresentados na Tabela 2, tendo sido obtidos da literatura.

As seguintes métricas de avaliação são adotadas: o Índice de Calinski-Harabasz (CH) [Caliński and Harabasz 1974], o Índice de Rand Corrigido (CR) [Hubert and Arabie 1985], o Índice de Davies-Bouldin (DB) [Davies and Bouldin 1979], o Índice de Jaccard [Halkidi et al. 2002] e o Índice de Rand [Rand 1971]. As métricas de avaliação representam quatro métricas de maximização (CH, CR, JI e RI, situação representada por  $\uparrow$ ), e uma métrica de minimização (DB, situação representada por  $\downarrow$ ). O CH foi adotado como função de *fitness* para todos os modelos.

A análise experimental inclui uma avaliação empírica baseada no valor médio obtido para cada uma das métricas adotadas em relação a trinta execuções independentes dos experimentos para cada base de dados, assim como um sistema de *ranks* elaborado através da aplicação de testes de hipóteses do tipo *Teste de Friedman* [Friedman 1937] aos resultados. O teste de Friedman é um teste de hipóteses não-paramétrico que calcula valores de *ranks* para os algoritmos para cada base de dados separadamente. Se a hipótese nula de que os *ranks* não são significativamente diferentes for rejeitada, o teste de Nemenyi [Nemenyi 1962] é adotado com um teste *post hoc* para o teste de Friedman. De acordo com o teste de Nemenyi, a performance de dois algoritmos é considerada significativamente diferente se a diferença entre seus valores médios de *rank* for ao menos maior que uma *diferença crítica* dada por:

$$CD = q_\alpha \sqrt{\frac{n_{alg}(n_{alg}+1)}{6n_{bases}}} \quad (6)$$

onde  $n_{bases}$  representa o número de bases de dados (ou seja, de imagens),  $n_{alg}$  representa o número de algoritmos comparados e  $q_\alpha$  são valores críticos baseados em estatísticas



nos limites do modelo  $t$  de *Student* divididas por  $\sqrt{2}$  [Demšar 2006]. Os melhores *ranks* apontados pelo teste de Friedman-Nemenyi apresentarão valores altos, para as métricas de maximização (CH, CR, JI e RI), e valores baixos para métricas de minimização (DB).

A Tabela 3 apresenta os resultados experimentais. Em uma análise empírica, os três modelos propostos foram capazes de atingir desempenhos mais estáveis que o GSO tradicional, além de terem sido mais estáveis que as demais abordagens na maioria dos problemas avaliados. Os modelos propostos foram capazes de encontrar os melhores valores para a função de *fitness* na maioria dos casos (8 entre 9 bases de dados). Os algoritmos híbridos propostos foram capazes de encontrar o número real de agrupamentos para seis dos nove casos avaliados, o que pode ser considerado um bom desempenho, uma vez que os demais algoritmos testados obtiveram desempenhos semelhantes, e que esses resultados também são compatíveis com outros trabalhos da literatura [Das et al. 2007, Tam et al. 2017]. Mesmo para os três casos nos quais não houve um casamento perfeito entre o valor esperado e o obtido, os valores estimados pelos modelos foram bastante próximos aos valores reais. A avaliação estatística obtida pelo sistema de *ranks* proporcionados pelos testes de Friedman-Nemenyi (Tabela 4) indica que os modelos híbridos propostos foram capazes de obter os melhores resultados globais, levando-se em conta cada uma das métricas de avaliação, tendo seus resultados sido considerados superiores, com diferenças estatísticas significativas em relação aos demais modelos (exceto para o Índice de Davies-Bouldin, no qual os resultados do RHGSO e BDHGSO foram considerados estatisticamente equivalentes aos do PSO e BGSO), inclusive quando comparados ao GSO original.

## 5. Conclusões

Neste trabalho, algoritmos híbridos para a tarefa de Agrupamento Automático de Dados são apresentados (RHGSO, ADHGSO e BDHGSO). As abordagens propostas combinam a boa capacidade de execução de buscas globais do GSO com a velocidade de exploração proporcionadas por heurísticas voltadas ao contexto de Agrupamento Automático de Dados. Os algoritmos propostos foram comparados a seis algoritmos do estado da arte em Computação Evolucionária e Inteligência de Enxames, adaptados ao contexto de Agrupamento Automático de Dados: GA, DE, PSO, BSA, GSO e BGSO. A análise experimental levou em consideração cinco métricas de agrupamentos de dados, assim como um teste de hipóteses estatístico, sendo realizada pela aplicação dos modelos selecionados a nove bases de dados *benchmark* reais. Os resultados experimentais relevaram que os modelos híbridos propostos apresentam maior estabilidade e boas performances que os algoritmos de comparação (inclusive GSO tradicional, e um modelo híbrido entre GSO e BSA), de acordo com uma avaliação estatística. Porém, a avaliação estatística indicou que, embora o modelo ADHGSO tenha obtido um *rank* global ligeiramente melhor que o RHGSO e o BDHGSO (considerando-se as cinco métricas avaliadas em conjunto), as três abordagens tiveram desempenhos estatisticamente equivalentes. Como trabalhos futuros, pretendemos estender nossa avaliação pela introdução de novos problemas à análise, tanto pela aquisição de bases de dados obtidas em cenários controlados (bases de dados sintéticas), quanto pela aplicação dos modelos propostos a problemas reais práticos, de modo a possibilitar uma investigação mais aprofundada em relação à cada uma das heurísticas propostas.

**Tabela 3. Resultados experimentais para as bases de dados reais (Média  $\pm$  Desvio Padrão).**

Base de Dados	Algoritmo	$CH^\dagger$	$CR^\dagger$	$DB^\ddagger$	$JI^\ddagger$	$RI^\ddagger$	$C$
Banknote Authentication	GA	1423.4 $\pm$ 0.2024	0.0487 $\pm$ 0.0015	0.8709 $\pm$ 0.0012	0.3803 $\pm$ 0.0008	0.5249 $\pm$ 0.0007	<b>2.000 <math>\pm</math> 0</b>
	DE	1423.6 $\pm$ 0.1534	0.0485 $\pm$ 0.0006	0.8704 $\pm$ 0.0009	0.3804 $\pm$ 0.0006	0.5249 $\pm$ 0.0003	<b>2.000 <math>\pm</math> 0</b>
	PSO	1387.5 $\pm$ 107.6	<b>0.0647 <math>\pm</math> 0.0420</b>	0.8863 $\pm$ 0.0378	0.3573 $\pm$ 0.0522	<b>0.5323 <math>\pm</math> 0.0198</b>	2.7667 $\pm$ 2.0625
	BSA	1423.5 $\pm$ 0.3278	0.0489 $\pm$ 0.0013	0.8707 $\pm$ 0.0009	0.3805 $\pm$ 0.0008	0.5251 $\pm$ 0.0007	<b>2.000 <math>\pm</math> 0</b>
	GSO	1423.6 $\pm$ 0.2776	0.0486 $\pm$ 0.0008	0.8701 $\pm$ 0.0004	0.3805 $\pm$ 0.0003	0.5249 $\pm$ 0.0004	<b>2.000 <math>\pm</math> 0</b>
	BGSO	<b>1423.7 <math>\pm</math> 0.0498</b>	0.0486 $\pm$ 0.0003	0.8702 $\pm$ 0.0004	0.3805 $\pm$ 0.0002	0.5249 $\pm$ 0.0001	<b>2.000 <math>\pm</math> 0</b>
	RHGSO	<b>1423.7 <math>\pm</math> 0.0111</b>	0.0486 $\pm$ 0.0002	0.8701 $\pm$ 0.0002	<b>0.3805 <math>\pm</math> 0.0001</b>	0.5249 $\pm$ 0.0001	<b>2.000 <math>\pm</math> 0</b>
	ADHGSO	<b>1423.7 <math>\pm</math> 0.0164</b>	0.0486 $\pm$ 0.0003	0.8701 $\pm$ 0.0002	0.3805 $\pm$ 0.0002	0.5249 $\pm$ 0.0002	<b>2.000 <math>\pm</math> 0</b>
	BDHGSO	<b>1423.7 <math>\pm</math> 0.0083</b>	0.0486 $\pm$ 0.0002	<b>0.8701 <math>\pm</math> 0.0001</b>	<b>0.3805 <math>\pm</math> 0.0001</b>	0.5249 $\pm$ 0.0001	<b>2.000 <math>\pm</math> 0</b>
Cancer	GA	1038.9 $\pm$ 1.9792	0.8320 $\pm$ 0.0090	0.7618 $\pm$ 0.0006	0.8599 $\pm$ 0.0067	0.9169 $\pm$ 0.0044	<b>2.000 <math>\pm</math> 0</b>
	DE	1038.9 $\pm$ 2.5719	0.8344 $\pm$ 0.0084	0.7618 $\pm$ 0.0006	0.8618 $\pm$ 0.0062	0.9181 $\pm$ 0.0041	<b>2.000 <math>\pm</math> 0</b>
	PSO	1029.3 $\pm$ 65.950	0.8372 $\pm$ 0.0121	0.7873 $\pm$ 0.1429	0.8633 $\pm$ 0.0116	0.9194 $\pm$ 0.0063	2.0333 $\pm$ 0.1826
	BSA	1038.9 $\pm$ 2.0758	0.8337 $\pm$ 0.0105	0.7618 $\pm$ 0.0006	0.8613 $\pm$ 0.0079	0.9177 $\pm$ 0.0052	<b>2.000 <math>\pm</math> 0</b>
	GSO	1041.3 $\pm$ 0.2277	<b>0.8396 <math>\pm</math> 0.0026</b>	<b>0.7612 <math>\pm</math> 0.0001</b>	<b>0.8655 <math>\pm</math> 0.0019</b>	<b>0.9206 <math>\pm</math> 0.0013</b>	<b>2.000 <math>\pm</math> 0</b>
	BGSO	<b>1041.4 <math>\pm</math> 0.0691</b>	0.8391 $\pm$ 0.0024	<b>0.7612 <math>\pm</math> 0.00005</b>	0.8651 $\pm$ 0.0018	0.9204 $\pm$ 0.0012	<b>2.000 <math>\pm</math> 0</b>
	RHGSO	<b>1041.4 <math>\pm</math> 0</b>	0.8391 $\pm$ 0	<b>0.7612 <math>\pm</math> 0</b>	0.8651 $\pm$ 0	0.9204 $\pm$ 0	<b>2.000 <math>\pm</math> 0</b>
	ADHGSO	<b>1041.4 <math>\pm</math> 0.1782</b>	0.8385 $\pm$ 0.0029	<b>0.7612 <math>\pm</math> 0.0001</b>	0.8647 $\pm$ 0.0022	0.9201 $\pm$ 0.0014	<b>2.000 <math>\pm</math> 0</b>
	BDHGSO	<b>1041.4 <math>\pm</math> 0.0540</b>	0.8392 $\pm$ 0.0010	<b>0.7612 <math>\pm</math> 0.00005</b>	0.8653 $\pm$ 0.0007	0.9204 $\pm$ 0.0005	<b>2.000 <math>\pm</math> 0</b>
Diabetes	GA	1139.1 $\pm$ 2.102	0.0443 $\pm$ 0.0036	0.6646 $\pm$ 0.0042	0.3789 $\pm$ 0.0041	0.5233 $\pm$ 0.0022	<b>3 <math>\pm</math> 0</b>
	DE	1140.0 $\pm$ 2.251	0.0450 $\pm$ 0.0025	0.6651 $\pm$ 0.0032	0.3793 $\pm$ 0.0026	0.5236 $\pm$ 0.0014	<b>3 <math>\pm</math> 0</b>
	PSO	996.17 $\pm$ 187.6	<b>0.0501 <math>\pm</math> 0.0164</b>	0.8084 $\pm$ 0.2226	0.3277 $\pm$ 0.0969	0.5188 $\pm$ 0.0161	4.6667 $\pm$ 2.928
	BSA	1136.5 $\pm$ 3.586	0.0453 $\pm$ 0.0046	<b>0.6638 <math>\pm</math> 0.0037</b>	<b>0.3806 <math>\pm</math> 0.0050</b>	<b>0.5242 <math>\pm</math> 0.0030</b>	<b>3 <math>\pm</math> 0</b>
	GSO	1141.8 $\pm$ 2.930	0.0451 $\pm$ 0.0010	0.6673 $\pm$ 0.0044	0.3783 $\pm$ 0.0017	0.5233 $\pm$ 0.0004	<b>3 <math>\pm</math> 0</b>
	BGSO	1142.1 $\pm$ 0.9429	0.0446 $\pm$ 0.0014	0.6679 $\pm$ 0.0018	0.3777 $\pm$ 0.0015	0.5232 $\pm$ 0.0008	<b>3 <math>\pm</math> 0</b>
	RHGSO	<b>1142.6 <math>\pm</math> 0.0240</b>	0.0451 $\pm$ 0.0003	0.6681 $\pm$ 0.00009	0.3781 $\pm$ 0.0002	0.5235 $\pm$ 0.0001	<b>3 <math>\pm</math> 0</b>
	ADHGSO	<b>1142.5 <math>\pm</math> 0.3924</b>	0.0452 $\pm$ 0.0004	0.6681 $\pm$ 0.0002	0.3782 $\pm$ 0.0005	0.5236 $\pm$ 0.0003	<b>3 <math>\pm</math> 0</b>
	BDHGSO	<b>1142.5 <math>\pm</math> 0.4049</b>	0.0452 $\pm$ 0.0003	0.6681 $\pm$ 0.0002	0.3781 $\pm$ 0.0002	0.5235 $\pm$ 0.0002	<b>3 <math>\pm</math> 0</b>
Heart	GA	<b>206.95 <math>\pm</math> 0.0036</b>	0.0295 $\pm$ 0.0012	0.9875 $\pm$ 0.0006	0.3606 $\pm$ 0.0009	0.5150 $\pm$ 0.0006	<b>2 <math>\pm</math> 0</b>
	DE	<b>206.95 <math>\pm</math> 0</b>	<b>0.0302 <math>\pm</math> 0</b>	<b>0.9871 <math>\pm</math> 0</b>	<b>0.3611 <math>\pm</math> 0</b>	<b>0.5154 <math>\pm</math> 0</b>	<b>2 <math>\pm</math> 0</b>
	PSO	206.84 $\pm$ 0.0995	0.0250 $\pm$ 0.0037	<b>0.9871 <math>\pm</math> 0.0014</b>	0.3591 $\pm$ 0.0012	0.5128 $\pm$ 0.0018	<b>2 <math>\pm</math> 0</b>
	BSA	<b>206.95 <math>\pm</math> 0</b>	<b>0.0302 <math>\pm</math> 0</b>	<b>0.9871 <math>\pm</math> 0</b>	<b>0.3611 <math>\pm</math> 0</b>	0.5153 $\pm$ 0.0003	<b>2 <math>\pm</math> 0</b>
	GSO	<b>206.95 <math>\pm</math> 0.0041</b>	0.0301 $\pm$ 0.0005	0.9873 $\pm$ 0.0007	0.3610 $\pm$ 0.0006	<b>0.5154 <math>\pm</math> 0</b>	<b>2 <math>\pm</math> 0</b>
	BGSO	<b>206.95 <math>\pm</math> 0.0015</b>	0.0301 $\pm$ 0.0005	0.9872 $\pm$ 0.0003	0.3610 $\pm$ 0.0004	0.5153 $\pm$ 0.0002	<b>2 <math>\pm</math> 0</b>
	RHGSO	<b>206.95 <math>\pm</math> 0</b>	<b>0.0302 <math>\pm</math> 0.00002</b>	<b>0.9871 <math>\pm</math> 0.0006</b>	<b>0.3611 <math>\pm</math> 0.0004</b>	<b>0.5154 <math>\pm</math> 0</b>	<b>2 <math>\pm</math> 0</b>
	ADHGSO	<b>206.95 <math>\pm</math> 0</b>	<b>0.0302 <math>\pm</math> 0</b>	<b>0.9871 <math>\pm</math> 0</b>	<b>0.3611 <math>\pm</math> 0</b>	<b>0.5154 <math>\pm</math> 0</b>	<b>2 <math>\pm</math> 0</b>
	BDHGSO	<b>206.95 <math>\pm</math> 0.0003</b>	<b>0.0302 <math>\pm</math> 0</b>	<b>0.9871 <math>\pm</math> 0</b>	<b>0.3611 <math>\pm</math> 0</b>	<b>0.5154 <math>\pm</math> 0</b>	<b>2 <math>\pm</math> 0</b>
Ionosphere	GA	115.65 $\pm$ 1.198	0.1464 $\pm$ 0.0132	1.5341 $\pm$ 0.0111	0.4190 $\pm$ 0.0064	0.5734 $\pm$ 0.0066	<b>2 <math>\pm</math> 0</b>
	DE	115.48 $\pm$ 1.601	0.1427 $\pm$ 0.0158	1.5367 $\pm$ 0.0143	0.4175 $\pm$ 0.0074	0.5716 $\pm$ 0.0078	<b>2 <math>\pm</math> 0</b>
	PSO	116.13 $\pm$ 9.484	<b>0.1791 <math>\pm</math> 0.0214</b>	1.5375 $\pm$ 0.0895	0.4317 $\pm$ 0.0084	<b>0.5893 <math>\pm</math> 0.0099</b>	2.0667 $\pm$ 0.258
	BSA	117.27 $\pm$ 0.9134	0.1564 $\pm$ 0.0151	1.5206 $\pm$ 0.0094	0.4233 $\pm$ 0.0075	0.5807 $\pm$ 0.0073	<b>2 <math>\pm</math> 0</b>
	GSO	118.43 $\pm$ 0.3889	0.1697 $\pm$ 0.0091	1.5158 $\pm$ 0.0052	0.4298 $\pm$ 0.0043	0.5845 $\pm$ 0.0055	<b>2 <math>\pm</math> 0</b>
	BGSO	118.60 $\pm$ 0.2825	0.1716 $\pm$ 0.0062	1.5137 $\pm$ 0.0035	0.4306 $\pm$ 0.0031	0.5859 $\pm$ 0.0031	<b>2 <math>\pm</math> 0</b>
	RHGSO	<b>118.83 <math>\pm</math> 0.0030</b>	0.1770 $\pm$ 0.0016	<b>1.5132 <math>\pm</math> 0.0007</b>	<b>0.4333 <math>\pm</math> 0.0009</b>	0.5886 $\pm$ 0.0008	<b>2 <math>\pm</math> 0</b>
	ADHGSO	118.80 $\pm$ 0.0848	0.1763 $\pm$ 0.0025	1.5133 $\pm$ 0.0014	0.4329 $\pm$ 0.0012	0.5883 $\pm$ 0.0016	<b>2 <math>\pm</math> 0</b>
	BDHGSO	118.81 $\pm$ 0.0649	0.1762 $\pm$ 0.0026	<b>1.5132 <math>\pm</math> 0.0014</b>	0.4329 $\pm$ 0.0013	0.5882 $\pm$ 0.0013	<b>2 <math>\pm</math> 0</b>
Iris	GA	561.58 $\pm$ 0.256	0.7302 $\pm$ 0.0001	0.6622 $\pm$ 0.0013	0.6958 $\pm$ 0.0003	0.8797 $\pm$ 0	<b>3 <math>\pm</math> 0</b>
	DE	<b>561.63 <math>\pm</math> 0</b>	0.7302 $\pm$ 0	<b>0.6620 <math>\pm</math> 0</b>	0.6959 $\pm$ 0	0.8797 $\pm$ 0	<b>3 <math>\pm</math> 0</b>
	PSO	560.80 $\pm$ 2.540	0.7301 $\pm$ 0.0004	0.6636 $\pm$ 0.0047	0.6956 $\pm$ 0.0007	0.8797 $\pm$ 0	<b>3 <math>\pm</math> 0</b>
	BSA	<b>561.63 <math>\pm</math> 0</b>	0.7302 $\pm$ 0	<b>0.6620 <math>\pm</math> 0</b>	0.6959 $\pm$ 0	0.8797 $\pm$ 0	<b>3 <math>\pm</math> 0</b>
	GSO	561.37 $\pm$ 0.8113	<b>0.7316 <math>\pm</math> 0.0040</b>	0.6627 $\pm$ 0.0023	<b>0.6971 <math>\pm</math> 0.0037</b>	<b>0.8803 <math>\pm</math> 0.0019</b>	<b>3 <math>\pm</math> 0</b>
	BGSO	<b>561.63 <math>\pm</math> 0</b>	0.7302 $\pm$ 0	<b>0.6620 <math>\pm</math> 0</b>	0.6959 $\pm$ 0	0.8797 $\pm$ 0	<b>3 <math>\pm</math> 0</b>
	RHGSO	<b>561.63 <math>\pm</math> 0</b>	0.7302 $\pm$ 0	<b>0.6620 <math>\pm</math> 0</b>	0.6959 $\pm$ 0	0.8797 $\pm$ 0	<b>3 <math>\pm</math> 0</b>
	ADHGSO	<b>561.63 <math>\pm</math> 0</b>	0.7302 $\pm$ 0	<b>0.6620 <math>\pm</math> 0</b>	0.6959 $\pm$ 0	0.8797 $\pm$ 0	<b>3 <math>\pm</math> 0</b>
	BDHGSO	<b>561.63 <math>\pm</math> 0</b>	0.7302 $\pm$ 0	<b>0.6620 <math>\pm</math> 0</b>	0.6959 $\pm$ 0	0.8797 $\pm$ 0	<b>3 <math>\pm</math> 0</b>
Page Blocks Classification	GA	14395.2 $\pm$ 567.3	0.0070 $\pm$ 0.0154	<b>0.5250 <math>\pm</math> 0.0342</b>	0.6044 $\pm$ 0.0893	0.6287 $\pm$ 0.0757	5.5 $\pm$ 0.509
	DE	<b>16343.2 <math>\pm</math> 778.1</b>	0.0003 $\pm$ 0.0129	0.6159 $\pm$ 0.0311	0.5195 $\pm$ 0.0745	0.5576 $\pm$ 0.0619	7.5 $\pm$ 0.861
	PSO	13372.5 $\pm$ 1920.9	0.0109 $\pm$ 0.0059	0.5307 $\pm$ 0.0318	0.6634 $\pm$ 0.0300	0.6782 $\pm$ 0.0260	<b>4.7000 <math>\pm</math> 0.8769</b>
	BSA	15007.2 $\pm$ 529.4	0.0057 $\pm$ 0.0156	0.5626 $\pm$ 0.0523	0.6031 $\pm$ 0.0735	0.6196 $\pm$ 0.0675	5.9667 $\pm$ 0.8087
	GSO	12456.9 $\pm$ 1436.9	0.0110 $\pm$ 0.0108	0.5364 $\pm$ 0.0267	0.6667 $\pm$ 0.0232	0.6874 $\pm$ 0.0548	4.3667 $\pm$ 0.5561
	BGSO	12583.7 $\pm$ 1668.9	0.0125 $\pm$ 0.0181	0.5306 $\pm$ 0.0262	0.6711 $\pm$ 0.0371	0.6849 $\pm$ 0.0333	4.4 $\pm$ 0.6747
	RHGSO	12273.9 $\pm$ 1183.0	0.0118 $\pm$ 0.0062	0.5496 $\pm$ 0.0286	0.6677 $\pm$ 0.0222	0.6818 $\pm$ 0.0193	4.4 $\pm$ 0.6215
	ADHGSO	12289.7 $\pm$ 1377.9	<b>0.0188 <math>\pm</math> 0.0183</b>	0.5277 $\pm$ 0.0282	<b>0.6830 <math>\pm</math> 0.0285</b>	<b>0.6957 <math>\pm</math> 0.0263</b>	4.2667 $\pm$ 0.5208
	BDHGSO	12123.1 $\pm$ 1264.8	0.0181 $\pm$ 0.0215	0.5375 $\pm$ 0.0184	0.6773 $\pm$ 0.0474	0.6909 $\pm$ 0.0423	4.2333 $\pm$ 0.6789
Seeds	GA	375.31 $\pm$ 0.7548	<b>0.7178 <math>\pm</math> 0.0086</b>	0.7535 $\pm$ 0.0010	<b>0.6827 <math>\pm</math> 0.0081</b>	<b>0.8749 <math>\pm</math> 0.0039</b>	<b>3 <math>\pm</math> 0</b>
	DE	372.38 $\pm$ 2.3840	0.7106 $\pm$ 0.0209	0.7564 $\pm$ 0.0041	0.6763 $\pm$ 0.0194	0.8716 $\pm$ 0.0094	<b>3 <math>\pm</math> 0</b>
	PSO	375.73 $\pm$ 0.2892	0.7159 $\pm$ 0.0028	0.7535 $\pm$ 0.0007	0.6808 $\pm$ 0.0026	0.8740 $\pm$ 0.0012	<b>3 <math>\pm</math> 0</b>
	BSA	370.66 $\pm$ 5.5973	0.6988 $\pm$ 0.0274	0.7603 $\pm$ 0.0081	0.6656 $\pm$ 0.0243	0.8627 $\pm$ 0.0261	<b>3 <math>\pm</math> 0</b>
	GSO	375.68 $\pm$ 0.3881	0.7153 $\pm$ 0.0040	0.7535 $\pm$ 0.0007	0.6803 $\pm$ 0.0037	0.8745 $\pm$ 0.0009	<b>3 <math>\pm</math> 0</b>
	BGSO	374.30 $\pm$ 4.9807	0.7095 $\pm$ 0.0452	0.7514 $\pm$ 0.0121	0.6772 $\pm$ 0.0300	0.8701 $\pm$ 0.0261	2.9667 $\pm$ 0.1826
	RHGSO	<b>375.81 <math>\pm</math> 0</b>	0.7166 $\pm$ 0	<b>0.7533 <math>\pm</math> 0</b>	0.6815 $\pm$ 0	0.8744 $\pm$ 0	<b>3 <math>\pm</math> 0</b>
	ADHGSO	<b>375.81 <math>\pm</math> 0</b>	0.7166 $\pm$ 0	<b>0.7533 <math>\pm</math> 0</b>	0.6815 $\pm$ 0	0.8744 $\pm$ 0	<b>3 <math>\pm</math> 0</b>
	BDHGSO	375.77 $\pm$ 0.2081	0.7163 $\pm$ 0.0020	0.7534 $\pm$ 0.0005	0.6812 $\pm$ 0.0019	0.8742 $\pm$ 0.0005	<b>3 <math>\pm</math> 0</b>
Waveform	GA	2518.7 $\pm$ 11.88					

**Tabela 4. Avaliação Global: *Ranks* médios para o Teste de Friedman-Nemenyi, com  $CD = 4.0043$ . O escore de cada algoritmo por métrica é apresentado (em parênteses).**

Algoritmo	$CH^t$	$CR^t$	$DB^t$	$JJ^t$	$RI^t$	Escore Médio
GA	86.0278 (9)	101.5481 (9)	161.6222 (7)	107.5037 (8)	101.8185 (9)	8.4
DE	111.9648 (7)	108.5444 (8)	172.1852 (9)	107.3667 (9)	104.1481 (8)	8.2
PSO	141.6685 (5)	142.5889 (4)	118.6463 (4)	136.2648 (6)	141.5870 (5)	4.8
BSA	89.5833 (8)	116.1519 (7)	168.6759 (8)	118.7463 (7)	116.1481 (7)	7.4
GSO	135.3741 (6)	141.1519 (5)	132.7759 (6)	138.7500 (5)	139.7352 (6)	5.6
BGSO	148.8315 (4)	140.0241 (6)	120.7611 (5)	143.8630 (4)	143.1444 (4)	4.6
RHGSO	<b>170.8685</b> (1)	<b>155.2704</b> (2)	<b>116.3759</b> (3)	<b>155.1519</b> (2)	<b>156.7111</b> (2)	<b>2.0</b>
ADHGSO	<b>167.8167</b> (2)	<b>158.6074</b> (1)	<b>113.2204</b> (1)	<b>157.9259</b> (1)	<b>159.7519</b> (1)	<b>1.2</b>
BDHGSO	<b>167.3648</b> (3)	<b>155.6130</b> (2)	<b>115.2370</b> (2)	<b>153.9278</b> (3)	<b>156.4556</b> (3)	<b>2.6</b>

## Agradecimentos

Os autores gostariam de agradecer ao CNPq, à CAPES e à FACEPE, agências de fomento brasileiras.

## Referências

- Asuncion, A. and Newman, D. (2007). Uci machine learning repository.
- Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.
- Civicioglu, P. (2013). Backtracking search optimization algorithm for numerical optimization problems. *Applied Mathematics and computation*, 219(15):8121–8144.
- Das, S., Abraham, A., and Konar, A. (2007). Automatic clustering using an improved differential evolution algorithm. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, 38(1):218–237.
- Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227.
- Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30.
- Dorigo, M., Maniezzo, V., and Colorni, A. (1996). Ant system: optimization by a colony of cooperating agents. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 26(1):29–41.
- Ezugwu, A. E., Shukla, A. K., Agbaje, M. B., Oyelade, O. N., José-García, A., and Agushaka, J. O. (2021). Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature. *Neural Computing and Applications*, 33(11):6247–6306.
- Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200):675–701.
- Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2002). Cluster validity methods: part i. *ACM Sigmod Record*, 31(2):40–45.
- He, S., Wu, Q. H., and Saunders, J. R. (2009). Group search optimizer: an optimization algorithm inspired by animal searching behavior. *IEEE Transactions on Evolutionary Computation*, 13(5):973–990.

- Holland, J. H. (1992). Genetic algorithms. *Scientific american*, 267(1):66–72.
- Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):193–218.
- Ikotun, A. M., Almutari, M. S., and Ezugwu, A. E. (2021). K-means-based nature-inspired metaheuristic algorithms for automatic data clustering problems: Recent advances and future directions. *Applied Sciences*, 11(23):11246.
- José-García, A. and Gómez-Flores, W. (2016). Automatic clustering using nature-inspired metaheuristics: A survey. *Applied Soft Computing*, 41:192–213.
- Kennedy, J. and Eberhart, R. (1995). Particle swarm optimization. In *Neural Networks, 1995. Proceedings., IEEE International Conference on*, volume 4, pages 1942–1948. IEEE.
- Nemenyi, P. (1962). Distribution-free multiple comparisons. In *Biometrics*, volume 18, page 263. INTERNATIONAL BIOMETRIC SOC 1441 I ST, NW, SUITE 700, WASHINGTON, DC 20005-2210.
- Pacífico, L. (2020). Agrupamento de imagens baseado em uma abordagem híbrida entre a otimização por busca em grupo e k-means para a segmentação automática de doenças em plantas. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, pages 152–163. SBC.
- Pacifico, L. and Ludermir, T. (2020). Backtracking group search optimization: A hybrid approach for automatic data clustering. In *Brazilian Conference on Intelligent Systems*, pages 64–78. Springer.
- Pacifico, L. D. and Ludermir, T. B. (2021). An evaluation of k-means as a local search operator in hybrid memetic group search optimization for data clustering. *Natural Computing*, 20(3):611–636.
- Preetha, V. (2021). Data analysis on student’s performance based on health status using genetic algorithm and clustering algorithms. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, pages 836–842. IEEE.
- Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.
- Shi, X., Zhang, X., and Xu, M. (2020). A self-adaptive preferred learning differential evolution algorithm for task scheduling in cloud computing. In *2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA)*, pages 145–148. IEEE.
- Storn, R. and Price, K. (1995). Differential evolution—a simple and efficient adaptive scheme for global optimization over continuous spaces. international computer science institute, berkeley. Technical report, CA, 1995, Tech. Rep. TR-95–012.
- Tam, H.-H., Ng, S.-C., Lui, A. K., and Leung, M.-F. (2017). Improved activation schema on automatic clustering using differential evolution algorithm. In *2017 IEEE Congress on Evolutionary Computation (CEC)*, pages 1749–1756. IEEE.
- Ye, L. and Zheng, D. (2021). Stable grasping control of robot based on particle swarm optimization. In *2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, pages 1020–1024. IEEE.