ĐẠI HỌC QUỐC GIA TP HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO MÔN HỌC MÁY HỌC

**Đề tài :**

**NGHIÊN CỨU VÀ SO SÁNH CÁC THUẬT TOÁN ĐOM ĐÓM LAI TRONG BÀI TOÁN PHÂN CỤM DỮ LIỆU TỰ ĐỘNG**

**Dựa trên bài báo :**

A Comparative Performance Study of Hybrid Firefly  Algoritms for Automatic Data Clustering

(Absalom el-shamir Ezugwu, Moyinoluwa b. Agbaje1, Nahla Aljojo,Rosanne Els,Haruna Chiroma, and Mohamed Abd Elaziz)

**GIẢNG VIÊN HƯỚNG DẪN :** TS. Trần Thái Sơn

**HỌC VIÊN THỰC HIỆN :**

20C12007 - Trần Đình Lâm

20C11040 - Đặng Nhật Minh

20C11035 - Trương Thế Kiệt

**KHÓA :** 30

**NGÀNH : Khoa học máy tính**

TP.HCM, 04-2021

**Mục Lục**

[**Mục Lục** 1](#_Toc70413272)

[**Bảng thông tin chi tiết nhóm** 1](#_Toc70413273)

[**Bảng chú thích từ viết tắt** 2](#_Toc70413274)

[**Danh sách bảng** 2](#_Toc70413275)

[**Danh sách hình ảnh minh họa** 2](#_Toc70413276)

[**1.** **Giới thiệu tổng quan** 3](#_Toc70413277)

[**1.1.** **Bài toán Clustering** 3](#_Toc70413278)

[**Khái niệm** 3](#_Toc70413279)

[**Phân loại các thuật toán Clustering** 3](#_Toc70413280)

[**1.2.** **Bài toán Optimization** 3](#_Toc70413281)

[**2.** **Cơ sở lý thuyết** 3](#_Toc70413282)

[**2.1.** **Thuật toán đom đóm FA** 3](#_Toc70413283)

[**2.2.** **Thuật toán PSO** 7](#_Toc70413284)

[**2.3.** **Thuật toán ABC** 11](#_Toc70413285)

[**2.4.** **Thuật toán IWO** 13](#_Toc70413286)

[**2.5.** **Thuật toán TLBO** 15](#_Toc70413287)

[**2.6.** **Lịch sử cải tiến, lai tạo giữa các thuật toán** 15](#_Toc70413288)

[**3.** **So sánh các phương pháp lai được bài báo đề xuất** 15](#_Toc70413289)

[**3.1.** **Các hàm mục tiêu được sử dụng** 15](#_Toc70413290)

[**a.** **Compact-Separated** 15](#_Toc70413291)

[**b.** **Davis-Bouldin** 15](#_Toc70413292)

[**3.2.** **4 Thuật toán lai được đề xuất** 15](#_Toc70413293)

[**4.** **Thực nghiệm và so sánh** 15](#_Toc70413294)

[**4.1.** **Dataset** 15](#_Toc70413295)

[**4.2.** **Cài đặt các bộ tham số phù hợp** 15](#_Toc70413296)

[**4.3.** **Kết quả thực nghiệm và phân tích** 15](#_Toc70413297)

[**4.3.1.** **Kết quả so sánh tổng quan** 15](#_Toc70413298)

[**4.3.2.** **Kết quả so sánh từng cặp thuật toán** 15](#_Toc70413299)

[**4.3.3.** **So sánh thời gian chạy thuật toán** 15](#_Toc70413300)

[**5.** **Demo** 15](#_Toc70413301)

[**6.** **Kết luận** 15](#_Toc70413302)

[References 16](#_Toc70413303)

# **Bảng thông tin chi tiết nhóm**

|  |  |  |  |
| --- | --- | --- | --- |
| **Mã nhóm:** | **CLA-05** | | |
| **Tên nhóm:** | **K2014** | | |
| **Số lượng:** | **3** | | |
| **MSSV** | **Họ tên** | **Email** | **Điện thoại** |
| 20C12007 | Trần Đình Lâm | tdlam123@gmail.com | 0383522356 |
| 20C11035 | Trương Thế Kiệt | truongthekiet709@gmail.com |  |
| 20C11040 | Đặng Nhật Minh | minhdangnhat685@gmail.com |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **BẢNG PHÂN CÔNG & ĐÁNH GIÁ HOÀN THÀNH CÔNG VIỆC** | | | |
| **Người thực hiện** | **Công việc thực hiện** | **Mức độ hoàn thành** | **Đánh giá của nhóm** |
| 20C12007  Trần Đình Lâm | Dịch bài báo sang tiếng Việt và giải thích các thuật ngữ | 90% | 9/10 |
| Chỉnh sửa slide |
|  |
| 20C11035  Trương Thế Kiệt | Thuật toán ABC | 80% | 8/10 |
| Bài toàn Optimization |
| Soạn slide cho buổi serminar |
| 20C11040  Đặng Nhật Minh | Thuật toán đom đóm, PSO | 80% | 8/10 |
|  |
| Soạn slide cho buổi serminar |

**Bảng chú thích từ viết tắt**

**Danh sách bảng**

**Danh sách hình ảnh minh họa**

1. **Giới thiệu tổng quan**
2. **Bài toán Clustering**

**Khái niệm**

**Phân loại các thuật toán Clustering**

1. **Bài toán Optimization**

Optimization là quá trình tìm giá trị tốt nhất(maximum/minimum) trong không gian dữ liệu lớn của một hàm mục tiêu bằng cách sử dụng các thuật toán. Trong đó việc tìm kiếm được giá tối ưu hóa được xem là vấn đề NP-Hard.

Trong bài toán Optimization, hàm mục tiêu là hàm cần tối ưu hóa trong bài toán. Hàm này được chọn, hoặc suy ra từ yêu cầu của bài toán. ta giải bài toán bằng cách tìm ra cực tiểu toàn cục (global optima) của hàm này, bằng các kỹ thuật, thuật toán hay phương pháp tối ưu hóa.

Những kỹ thuật tối ưu hóa thường được dùng đề tìm kiếm các giải pháp cho bài toán tối ưu hóa toàn cục. Trong đó chia thành 2 loại kỹ thuật khác nhau: Thuật toán chính xác và thuật toán gần đúng. Thuật toán chính xác là các mô hình toán học đã được dùng để tìm kiếm giải pháp tối ưu. Thuật toán gần đúng (metaheuristic) là giải pháp gần tối ưu trong một thời gian giải thuật ngắn, phù hợp cho những công trình phức tạp.

Thuật toán metaheuristic thường có thể tìm được lời giải gần tối ưu trong một khoảng thời gian ít hơn hơn, tốn ít công sức hơn. Thành phần chính của những phương pháp metaheuristic là thăm dò(exploration)- quá trình khám phá không gian tìm kiếm và tránh bẫy trong các cực trị địa phương(local optimal), khai thắc(exploiation)-Quá trình cải thiện bất kỳ giải pháp hứa hẹn nào đã thu được cho đến thời điểm hiện tại.

1. **Cơ sở lý thuyết**
2. **Thuật toán đom đóm FA**

Thuật toán đom đóm (Firefly algorithm - FA) là một thuật toán tối ưu hóa lấy cảm hứng từ thiên nhiên, dựa trên trí thông minh bầy đàn được phát triển bởi Xin-She Yang (2009) bắt chước cách bầy đom đóm giao tiếp với nhau

Đom đóm là một loại côn trùng có cánh và có thể phát ra ánh sáng nhấp nháy về đêm. Ánh sáng này dùng để thu hút các con đom đóm khác di chuyển về phía nó, và cảnh báo về các mối nguy hiểm hoặc nguồn thức ăn

**Ý tưởng thuật toán**

Ở thời điểm ban đầu mỗi con đom đóm trong bầy sẽ ở một vị trí ngẫu nhiên trong không gian tìm kiếm và di chuyển theo một hướng ngẫu nhiên. Khi một con đom đóm tìm thấy thức ăn, nó sẽ phát ra ánh sáng với cường độ tỉ lệ với lượng thức ăn tìm được. Mỗi con đóm đóm sẽ dựa vào ánh sáng để di chuyển về nơi có nhiều thức ăn nhất

**Nội dung thuật toán**

Cường độ ánh sáng trên đường đi sẽ giảm dần do bị hấp thụ một phần bởi môi trường, cho cường độ ánh sáng I tại điểm cách O một khoảng r, I tỉ lệ nghịch với bình phương khoảng cách:

với r là khoảng cách, I là cường độ ánh sáng tại điểm cách một khoảng r

Cường độ ánh sáng của mỗi con đom đóm phát ra sẽ tỉ lệ với hàm cần tối ưu của bài toán.

Thuật toán đom đóm được phát triển dựa trên giả thuyết rằng:

* Khi một con đom đóm (A) bị thu hút bởi một con con đom đóm khác (B) thì nó chỉ dựa trên độ sáng (mà nó cảm nhận được) của con đom đóm B và không bị ảnh hưởng bởi các yếu tố khác (giới tính đom đóm, mùi hương, mùa, ...)
* Độ thu hút tỉ lệ thuận với cường độ ánh sáng phát ra, một con đom đóm ít phát sáng hơn sẽ bị thu hút bởi một con đom đóm phát sáng sáng hơn, độ thu hút sẽ giảm dần theo khoảng cách giữa hai con đom đóm
* Nếu hai con đom đóm có cùng độ sáng (và là độ sáng lớn nhất trong quần thể) thì chúng sẽ di chuyển ngẫu nhiên

Độ sáng cảm nhận được tại 1 điểm và độ thu hút là 2 yếu tố quan trọng trong thuật toán đom đóm. Cường độ ánh sáng do con đom đóm phát ra tỉ lệ nghịch với giá trị của hàm mục tiêu tại điểm mà con đom đóm đang đứng, và trong quá trình truyền đi, độ sáng sẽ thay đổi theo độ lớn khoảng cách và lượng ánh sáng bị hấp thụ bởi môi trường, kết hợp lại, ta có phương trình:

với I0 là cường độ ánh sáng tại điểm cớ khoảng cách r = 0, gamma là hệ số hấp thụ ánh sáng của môi trường, r là khoảng cách

Vì độ thu hút tỉ lệ với cường độ ánh sáng cảm nhận được, ta có phương trình:

với β0­ là độ thu hút tại điểm r = 0

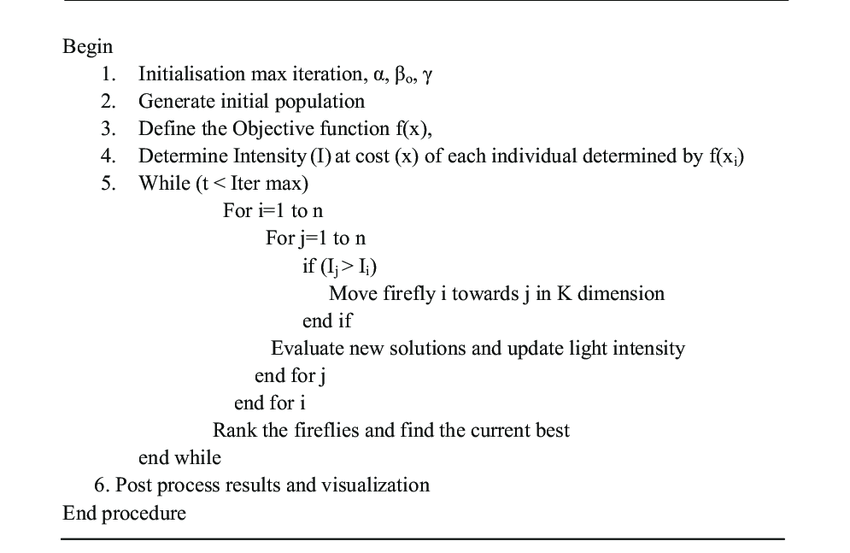
Khoảng cách Euclid giữa 2 con đom đóm xi và xj được tính bởi công thức:

với d là số chiều của dữ liệu

Sự chuyển động của 1 con đom đóm được xác định bởi công thức:

với , , là một số ngẫu nhiên từ phân phối Gauss

**Mã giả của thuật toán đom đóm**

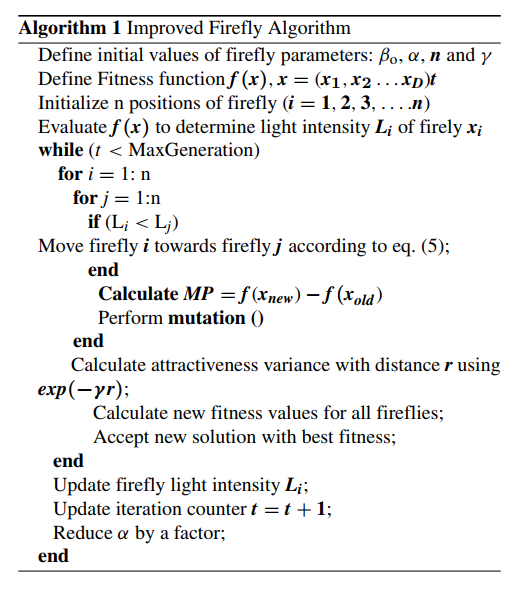


*Nhận xét: trong thuật toán ta thấy mỗi con đom đóm chỉ đi theo những con đom đóm sáng hơn nó mà không quan tâm tới personal best của mình, cách làm này sẽ giúp thuật toán hội tụ nhanh hơn, nhưng sẽ giảm khả năng tìm ra điểm cực trị tốt nhất*

Trong bài báo, để tăng khả năng khai thác và khám phá của thuật toán đom đóm, để có thể áp dụng vào những bài toán nhiều chiều một cách hiệu quả hơn, nhóm tác giả đã cải tiến thuật toán đom đóm bằng cách giới thiệu thuật ngữ mutation strategy (MP) để làm tăng độ thu hút của những con đom đóm có cường độ ánh sáng thấp hơn.

*Ý tưởng cơ bản là sau mỗi lần lặp, những con đom đóm bị đột biến sẽ di chuyển ngẫu nhiên thêm một khoảng nữa (để tăng khả năng tìm kiếm)*

Thuật toán đom đóm sau khi được cải tiến sẽ có dạng như sau:



1. **Thuật toán PSO**

Particle Swarm Optimazation (PSO) là một kĩ thuật tối ưu ngẫu nhiên dựa trên trí thông minh bầy đàn (population based stochastic optimization technique) được giới thiệu bởi Eberhart và Kennedy vào năm 1995 lấy cảm hứng từ cách di chuyển và trao đổi thông tin khi kiếm ăn của các thành viên trong quần thể bầy chim và đàn cá.

**Ý tưởng thuật toán**

Tại thời điểm ban đầu, mỗi thành viên trong quần thể ở một vị trí ngẫu nhiên và bay theo một hướng ngẫu nhiên. Sau một thời gian, một số thành viên tìm được thức ăn, và gửi tín hiệu tới các thành viên khác trong bầy tùy theo lượng thức ăn tìm thấy được. *Dựa vào thông tin đó, những thành viên khác sẽ di chuyển về nơi có nhiều thức ăn nhất (mà nó biết được).*

Ta có thể kết hợp PSO với một số thuật toán khác để giải quyết các bài toán về tối ưu.

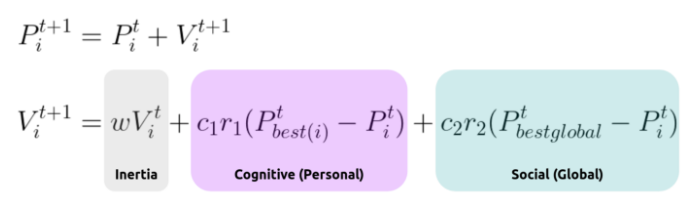
**Nội dung thuật toán**

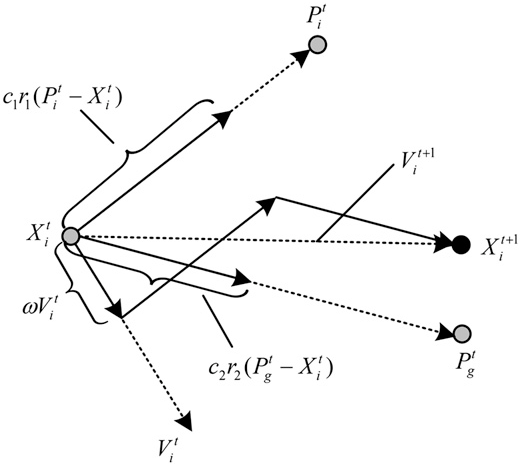
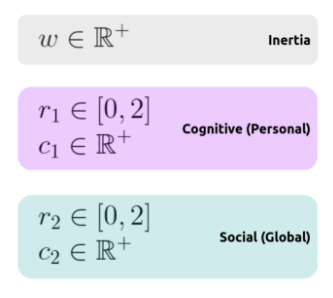
**Cá thể**: mỗi cá thể được xem như một giải pháp tiềm năng để đạt được global minimum trong không gian tìm kiếm *(dựa trên vị trí của nó)*. *Không con nào biết được chính xác global minimum ở đâu, nhưng dựa vào vị trí của mỗi con ta có thể tính được giá trị của hàm mục tiêu cần tối ưu*

Ở lúc ban đầu, mỗi cá thể sẽ được khởi tạo vị trí và vận tốc ngẫu nhiên.

với t là số thứ tự của lần lặp, i là số thứ tự của cá thể trong quần thể

**Quần thể**: qua mỗi lần lặp, vector vận tốc của mỗi cá thể được thay đổi một cách ngẫu nhiên dựa theo hướng của vị trí nhiều thức ăn nhất của cá thể đó tìm được (personal best), và hướng của vị trí có nhiều thức ăn nhất của quần thể (global best)





trong đó, các trọng số có ý nghĩa

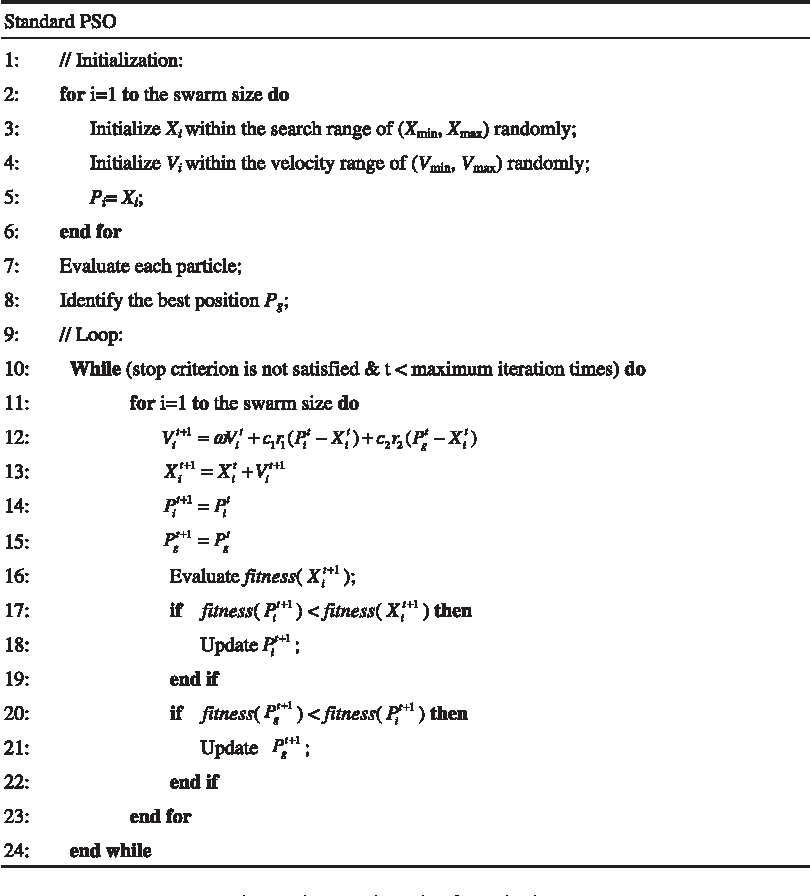
r1 và r2 là duy nhất cho mỗi cá thể và mỗi lần lặp  
c1 và c2 là hệ số cho biết độ phụ thuộc của vận tốc mới vào personal best và global best  
w là hệ số ứng với vận tốc hiện tại  
 là điểm personal best của cá thể thứ i tại lần lặp thứ t, là điểm global best tại lần lặp thứ t

*Bộ trọng số đó sẽ quyết định vận tốc (và vị trí) mới của cá thể, tạo ra sự cân bằng giữa khai phá và tìm kiếm (exploitaion and exploration) để tăng tốc hội tụ và tìm ra điểm cực trị tốt hơn*

*Nhận xét các trọng số:*

* *Với w lớn và c1, c2 nhỏ: quần thể khó hội tụ do mỗi cá thể chỉ đi theo hướng hiện tại của mình.*
* *Với c1 lớn và c2 nhỏ: quần thể hội tụ chậm do mỗi cá thể tập trung phần lớn vào hướng personal best của nó.*
* *Với c1 nhỏ và c2 lớn: quần thể hội tụ nhanh (do các cá thể đều hướng tới điểm global best của quần thể) nhưng điểm global minimum tìm được có thể không phải là điểm tốt nhất (do khả năng tìm kiếm bị hạn chế)*

**Mã giả thuật toán PSO**



1. **Thuật toán ABC**

Artificial Bee Colony (ABC) algorithm là thuật toán meta-heuristic dựa trên bầy đàn được giới thiệu năm 2005 bởi Karaboga để tối ưu các bài toán số.

**Ý tưởng thuật toán**

ACB lấy cảm hứng từ hành vi tìm kiếm thức ăn thông minh của ông mật.Trong đó bầy ông nhân tạo được chia thành ba nhóm: ong thợ (employed bees), ong giám sát(onlooker bees), ong trinh sát(scout bees).

* Ong thợ: đi đến nguồn thức ăn mà nó khám phá được trước đó.
* Ong quan sát: quan sát điệu nhảy của ong thợ để quyết định.
* Ong trinh sát: thực hiện tìm kiếm ngẫu nhiên để tìm kiếm nguồn thức ăn mới.

**Nội dung**

Thuật toán được chia làm bốn giai đoạn :

***Initialization Phase***

Khởi tạo ngẫu nhiên vị trí các food source, số lượng food source, số lượng ong thợ, ông giám sát, số lần chạy. Trong đó số lượng ong thợ, ông giám sát và food source là bằng nhau.

Tính toán giá trị của hàm mục tiêu, hàm fit(công thức tính giá trị hàm fit sẽ được giải thích ở phase Employed Bee).

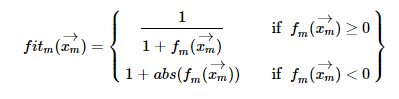
***Employed Bees Phase***

Tìm kiếm vị trị nguồn thức ăn lân cận dựa vào thông tin nguồn thức ăn trong kí ức trước đó. Vị trí nguồn thức ăn lân cận được tính theo công thức:

F:\SĐH\Kì_1\MAY_HOC\document\MayHoc\photos\abc_newfoodsourceposition.PNG

Sau khi tìm vị trí của nguồn thức ăn lân cần, các giá trị của hàm mục tiêu, hàm fit được tính toán lại. Nếu giá trị giá trị của hàm fit tại vị trí mới tốt hơn vị trí cũ trước đó, giá trị vị trí tốt nhất trong kí ức sẽ được cập nhật. Với mỗi lần mà vị trí nguồn thức ăn lân cận mới không tốt hơn vị trí cũ, số lần trial sẽ được tăng lên, và được reset khi giá trị tốt hơn.

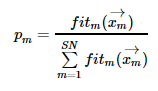
Giá trị của hàm fit sẽ được tính theo công thức sau:



Trong đó f(xm) là giá trị hàm mục tiêu tại vị trí xm

***Onlooker Bees Phase***

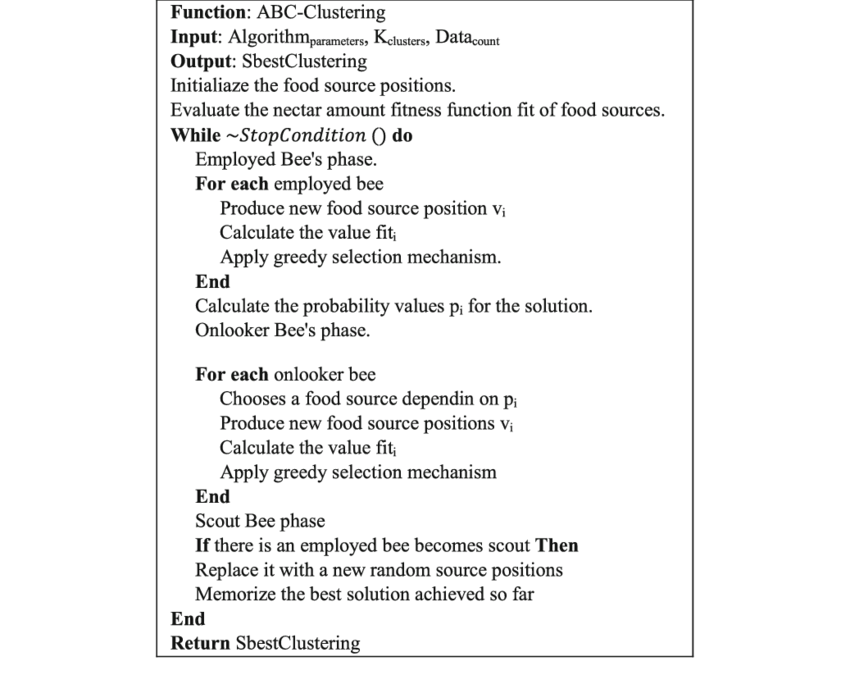
Quyết định lựa chọn nguồn thức ăn dựa vào giá trị xác suất được tính toán thông qua công thức:



***Scout Bees Phase***

Những con ông thợ khi giá trị trial đạt ngưỡng giới hạn (tức sau một số lần thức nghiệm tìm nguồn thức ăn lân cận nó không tìm được giải pháp tốt hơn) sẽ phải tìm lại vị trí nguồn thức ăn một cách ngẫu nhiên.

**Mã giã thuật toán ABC**



*Nhận xét:* *Nhìn chung thuật toán ABC có thể giúp hạn chế được các bẫy local optimal do sau một số lần giới hạn nếu không có giá trị khả thi hơn thì vị trí thức ăn tức giá trị của hàm mục tiêu sẽ được random lại và tiếp tục tìm kiếm.*

1. **Thuật toán IWO**

Invasive weed optimization(IWO): thuật toán tối ưu cỏ dại xâm lấn, lấy cảm hứng từ các quần thể cỏ dại ngoài tự nhiên dựa vào hành vi sống, sự thích nghi tốt với môi trường được đề xuất bởi Alireza Mehrabian và Caro Lucas năm 2006.

**Nội dung**

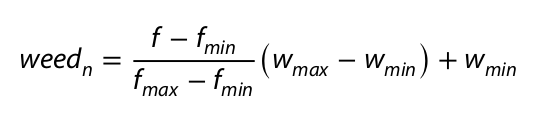
Thuật toán chia làm 4 giai đoạn:

* Initialization: Khởi tạo quần thể ngẫu nhiên các hạt trên không gian tìm kiếm
* Reproduction: Các hạt phát triển thành cây, và sinh ra nhiều hạt, tùy thuộc vào giá trị hiện tại
* Spatial dispersal: Các hạt sinh ra được tiếp tục phân tán, rồi phát triển thành cây mang giá trị
* Competitive exclusion: Khi đạt được số lượng cây quy định, tiến hành chọn lọc ra các cây có giá trị cao (low fitness), bỏ đi các cây khác kém hơn.

Quá trình loại trừ theo cơ chế sau:

* Mỗi cá thể đều sinh ra hạt dựa vào giá trị hiện tại của chính nó
* Các hạt sinh ra được phân tán ngẫu nhiên
* Giá trị của mỗi cá thể được so sánh với cha của nó, và chỉ giữ lại các cây có giá trị cao.

Công thức sinh hạt như sau:



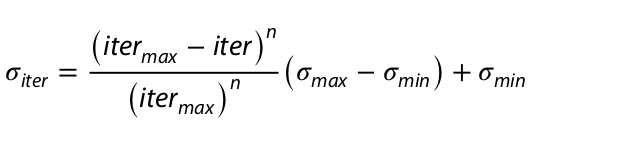
Trong đó:

f: giá trị sức sống hiện tại của cá thể

fmax và f\_min là 2 ngưỡng sức sống của quần thể hiện tại

wmax và wmin là 2 ngưỡng số lượng cá thể đã đặt ra

Quá trình phân tán hạt giống mô tả trong công thức:



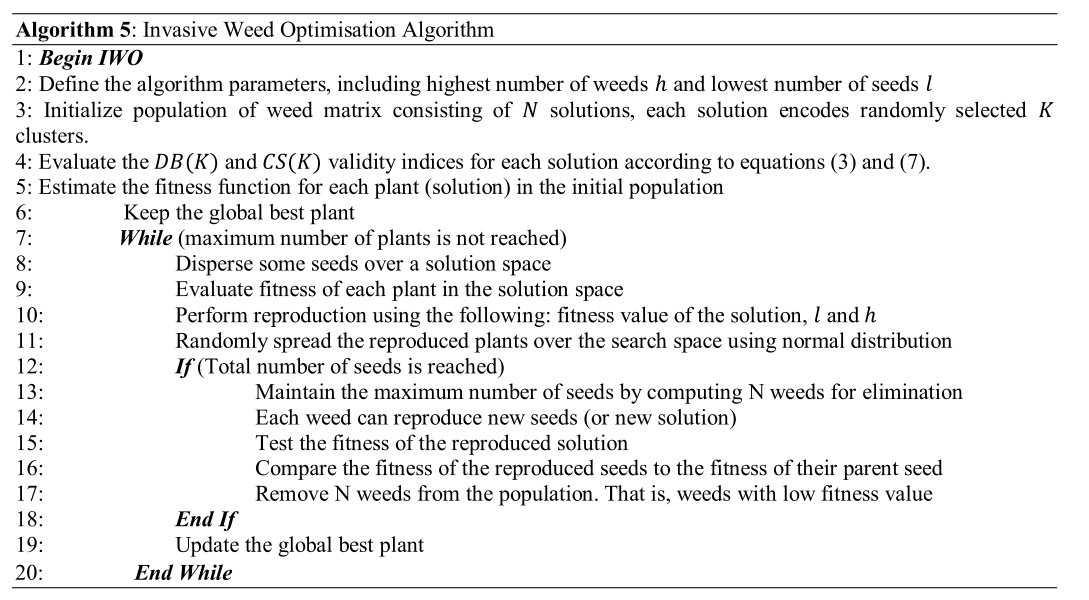
Trong đó:

Itermax là số vòng lặp tối đa

σiter là độ lệch chuẩn hiện tại

n là hệ số điều chỉnh

**Mã giả thuật toán**



1. **Thuật toán TLBO**
2. **Lịch sử cải tiến, lai tạo giữa các thuật toán**
3. **So sánh các phương pháp lai được bài báo đề xuất**
4. **Các hàm mục tiêu được sử dụng**
   1. **Compact-Separated**
   2. **Davis-Bouldin**
5. **4 Thuật toán lai được đề xuất**
6. **Thực nghiệm và so sánh**
7. **Dataset**
8. **Cài đặt các bộ tham số phù hợp**
9. **Kết quả thực nghiệm và phân tích**
10. **Kết quả so sánh tổng quan**
11. **Kết quả so sánh từng cặp thuật toán**
12. **So sánh thời gian chạy thuật toán**
13. **Demo**
14. **Kết luận**

Trong nghiên cứu này, nhóm tác giả đã kết hợp thuật toán đom đóm (FA) với 4 thuật toán khác bao gồm PSO, ABC, IWO và TLBO.

Kết quả thực nghiệm cho thấy: thuật toán lai giữa FA và PSO (FAPSO) cho kết quả vượt trội so với 3 thuật toán lại còn lại về chất lượng phân cụm và tốc độ chạy. Mặt khác thì FATLBO đạt được kết quả tốt thứ hai (sau FAPSO) nhưng có thời gian chạy nhanh hơn, FAIWO cho kết quả tệ nhất cả về mặt kết quả và thời gian chạy.

Nhóm tác giả định hướng tương lai sẽ kết hợp thuật toán đom đóm với các thuật toán khác để giải những bài toán phân cụm phức tạp hơn.

# References

* Absalom El-shamir Ezugwu 1, (Member, IEEE), Moyinoluwa B. Agbaje1, Nahla Aljojo 2, Rosanne Els 1, Haruna Chiroma 3, (Member, IEEE), And Mohamed Abd Elaziz (2020), A Comparative Performance Study of Hybrid Firefly Algorithms for Automatic Data Clustering
* Mehrabian AR, Lucas C (2006) A novel numerical optimization algorithm inspired from weed colonization. Ecol Inform 1(4):355–366