

Ứng dụng thuật toán học máy Self-Organizing Map để phân cụm thạch học

Phạm Sơn Tùng^{1,2,*}, Trần Viết Việt Anh^{1,2}



Use your smartphone to scan this QR code and download this article

¹Bộ môn Khoan và Khai thác Dầu khí, Khoa Kỹ thuật Địa chất và Dầu khí, Trường Đại học Bách Khoa TP. HCM

²Đại học Quốc gia Thành phố Hồ Chí Minh

Liên hệ

Phạm Sơn Tùng, Bộ môn Khoan và Khai thác Dầu khí, Khoa Kỹ thuật Địa chất và Dầu khí, Trường Đại học Bách Khoa TP. HCM

Đại học Quốc gia Thành phố Hồ Chí Minh

Email: phamsontung@hcmut.edu.vn

Lịch sử

- Ngày nhận: 17-11-2025
- Ngày sửa đổi: 30-01-2026
- Ngày chấp nhận: 03-06-2026
- Ngày đăng: 10-09-2026

DOI : <https://doi.org/10.32508/vnuhcmj-et.v9i3.1557>



Bản quyền

© Tạp chí ĐHQG TP.HCM. Đây là bài báo công bố mở được phát hành theo các điều khoản của the Creative Commons Attribution 4.0 International license.

TÓM TẮT

Trong bối cảnh ngành dầu khí ngày càng chú trọng đến việc tối ưu hóa quá trình phân tích và diễn giải dữ liệu địa chất nhằm nâng cao hiệu quả thăm dò và khai thác, các phương pháp học máy (machine learning) đã và đang khẳng định vai trò quan trọng trong việc xử lý khối lượng lớn dữ liệu địa vật lý có tính đa chiều và phức tạp. Sự phát triển của các kỹ thuật học không giám sát đã mở ra nhiều hướng tiếp cận mới trong việc khai thác tri thức tiềm ẩn từ dữ liệu giếng khoan mà không phụ thuộc hoàn toàn vào kinh nghiệm chủ quan của chuyên gia. Trong số đó, thuật toán Self-Organizing Map (SOM) được đánh giá là một công cụ hiệu quả nhờ khả năng thực hiện phân cụm đồng thời bảo toàn mối quan hệ phi tuyến giữa các biến đầu vào và trực quan hóa dữ liệu đa chiều dưới dạng bản đồ hai chiều dễ diễn giải.

SOM đã được ứng dụng rộng rãi trong nhiều lĩnh vực của ngành dầu khí, bao gồm phân loại thạch học, nhận dạng môi trường trầm tích, phân chia đơn vị dòng chảy thủy lực, dự đoán đặc tính tầng chứa, phân tích dữ liệu địa vật lý logs và hỗ trợ xây dựng mô hình vỉa. Thông qua việc biểu diễn cấu trúc phân bố của dữ liệu trên không gian bản đồ, thuật toán này giúp các nhà địa chất và kỹ sư mô nhận diện được các mẫu đặc trưng, làm rõ mối liên hệ giữa các tham số địa vật lý và hỗ trợ quá trình ra quyết định trong điều kiện dữ liệu phức tạp và không chắc chắn.

Bài báo này trình bày kết quả nghiên cứu ứng dụng thuật toán Self-Organizing Map trong phân cụm dữ liệu logs giếng khoan nhằm xác định các nhóm thạch học có tiềm năng. Dữ liệu đầu vào bao gồm các đường cong địa vật lý đặc trưng của giếng khoan, được tiền xử lý và chuẩn hóa trước khi đưa vào mô hình SOM để huấn luyện và phân tích. Kết quả nghiên cứu cho thấy SOM có khả năng phân tách hiệu quả các nhóm dữ liệu có đặc điểm tương đồng, phản ánh được tính dị thể của tầng chứa và hỗ trợ nhận diện các khoảng có tiềm năng chứa hydrocarbon. Những kết quả đạt được đã chứng minh tính khả thi và hiệu quả của SOM trong xử lý dữ liệu địa vật lý, đồng thời cung cấp thêm cơ sở khoa học phục vụ công tác dự báo, đánh giá tiềm năng giếng và nâng cao độ tin cậy trong quá trình thăm dò và khai thác dầu khí.

Từ khóa: Bản đồ tự tổ chức, phân cụm thạch học, học máy, mạng nơ-ron nhân tạo, địa vật lý

GIỚI THIỆU

Bản đồ tự tổ chức (Self-Organizing Map – SOM) là một dạng mạng nơ-ron do Teuvo Kohonen phát triển vào đầu những năm 1980. SOM thuộc nhóm thuật toán học không giám sát, nơi mà hệ thống tự học cách sắp xếp và biểu diễn dữ liệu mà không cần đến thông tin nhãn trước. Điều đặc biệt ở SOM là khả năng ánh xạ các dữ liệu đa chiều phức tạp về một không gian hai chiều hoặc ba chiều trực quan hơn, đồng thời vẫn bảo toàn được mối quan hệ hình học giữa các điểm dữ liệu gốc. Trong đó, SOM được xây dựng như một mô hình mạng nơ-ron có khả năng học cạnh tranh, trong đó các nơ-ron được sắp xếp thành một lưới hai chiều và sẽ tự điều chỉnh trọng số để biểu diễn cấu trúc của không gian dữ liệu đầu vào¹.

Việc phân cụm tướng đá (facies classification) truyền thống chủ yếu thường phụ thuộc vào kinh nghiệm chuyên gia và các phương pháp thống kê cổ điển, vốn

đễ bị sai lệch khi dữ liệu có tính phi tuyến mạnh hoặc bị nhiễu. Nhưng bản đồ SOM, với khả năng tự tổ chức và bảo toàn cấu trúc lân cận của dữ liệu, cung cấp một giải pháp trực quan, hiệu quả và linh hoạt hơn trong việc nhận diện các kiểu mẫu trong dữ liệu đầu vào từ dữ liệu logs như đường cong gamma ray, mật độ, sóng âm,... Từ bảng dữ liệu ghi nhật ký này, các loại đá ở độ sâu cụ thể sẽ được phân loại thành các loại đá facies khác nhau.

Quá trình minh giải thông thường sẽ tốn rất nhiều thời gian và công sức. SOM chắc chắn là một giải pháp cho vấn đề này vì nó loại bỏ việc minh giải thủ công tốn thời gian. SOM sẽ dễ dàng phân loại các khoảng đá thành các loại đá facies cụ thể cùng một lúc².

Bên cạnh đó, việc kết hợp SOM với phương pháp phân cụm khác như phân cụm phân cấp (hierarchical clustering) góp phần nâng cao chất lượng phân tích, từ đó mang lại cái nhìn rõ ràng ở những vùng dữ liệu chính

Trích dẫn bài báo này: Tùng P S, Anh T V V. **Ứng dụng thuật toán học máy Self-Organizing Map để phân cụm thạch học.** VNUHCM J. Eng. Technol. 2026; 9(3): 3294-3301.

xác hơn về cấu trúc địa chất dưới lòng đất, yếu tố then chốt trong việc đánh giá và khai thác tài nguyên dầu khí³.

Hiện nay, nhu cầu tích hợp các kỹ thuật học máy vào mô hình mô phỏng địa tầng và địa chất số đang ngày càng tăng. Thuật toán SOM có thể đóng vai trò là cầu nối giữa dữ liệu thực nghiệm (được lấy từ giếng khoan) và mô hình hóa địa chất bằng cách hỗ trợ phân nhóm dữ liệu một cách khách quan và có căn cứ⁴⁻⁶. Do đó, việc nghiên cứu và áp dụng Self-Organizing Map trong bài toán phân cụm là có ý nghĩa về mặt học thuật và giá trị thực tiễn cao trong hoạt động thăm dò và khai thác dầu khí hiện đại.

PHƯƠNG PHÁP NGHIÊN CỨU

Bản đồ tự tổ chức (Self Organizing Map)

Thuật toán SOM là một tập con của mạng nơ-ron nhân tạo (ANN), thực hiện dựa trên thuật toán học không giám sát, thuộc nhóm nơ-ron học cạnh tranh được sử dụng để huấn luyện dữ liệu đầu vào. Thông thường mục đích chính của SOM là tự động phân cụm và giảm chiều dữ liệu nhưng vẫn giữ toàn vẹn cấu trúc dữ liệu. Quá trình này diễn ra trong hai giai đoạn chính:

Giai đoạn 1: SOM được huấn luyện trên một tập dữ liệu đầu vào V cụ thể mà không cần nhãn, sau đó qua mỗi lần huấn luyện được cập nhật trọng số W để có sự liên kết giữa các điểm dữ liệu tương đồng về kích thước và tính chất.

Giai đoạn 2: Bản đồ đã huấn luyện có thể được hiệu chỉnh hoặc phân cụm dữ liệu đầu vào để phân loại thành các nhóm có cùng đặc tính (ví dụ: các lớp địa chất thạch học).

Quá trình huấn luyện SOM

Đầu tiên, chúng ta lấy một tập hợp các vectơ đầu vào (dữ liệu) và trình bày nó trên một lưới bản đồ 'n' chiều. Đối với hình vuông (Square SOMs) và hình lục giác (Hexagonal SOMs) ở dưới dạng 2D được tính bằng công thức sau :

$$\text{Tổng số nút} = \text{Tổng số chiều}^2$$

Với hình cầu (Spherical SOMs) trên hình học geodesic domes⁷:

$$N = f^2 * 10 + 2 \quad (01)$$

Trong đó:

- f : là tần số chia nhỏ tam giác
- N : số đỉnh tam giác

Một hình đa diện đều icosahedron có:

- 20 mặt (tam giác).

- 12 đỉnh.
- 30 cạnh.

Với $f = 1$

Số mặt tam giác:

$$M = 20 * f^2 \quad (02)$$

Công thức xác định số cạnh trong mọi đa diện đều: với N (đỉnh), M (mặt), C (cạnh):

$$N + C = M + 2 \quad (03)$$

Với mỗi nút trong bản đồ đều có 1 giá trị trọng số cho vectơ đầu vào khác nhau. Trong quá trình huấn luyện, mỗi trọng số tương ứng với mỗi nút sẽ có giá trị ngẫu nhiên và kết quả của bản đồ đã huấn luyện sẽ khác nhau.

Đối với mỗi mức dữ liệu đầu vào, "Best Matching Unit (BMU)" được tính toán. Giá trị này được tính bằng khoảng cách Euclidean giữa vectơ trọng số của mỗi nút và vectơ đầu vào đã chọn như sau:

$$Distance = \sqrt{\sum_{i=0}^{i=n} (V_i - W_i)^2} \quad (04)$$

Khoảng cách Euclid được sử dụng làm tiêu chuẩn để so sánh. Giá trị khoảng cách ngắn nhất sẽ được chọn là giá trị BMU. Sau quá trình tính giá trị BMU, sẽ xác định bán kính vùng lân cận thì theo thông thường để chọn một vùng bán kính lân cận ban đầu để cập nhật trọng số xung quanh nó, chọn bán kính ban đầu bằng công thức:

Nơ-ron được chọn là BMU sẽ được cập nhật trọng số. Trọng số của BMU sẽ được cập nhật theo công thức sau:

$$W' = W_0 + \theta_0 L_0 (V_0 - W_0) \quad (06)$$

Với L_0 là tốc độ học ban đầu (Initial Learning rate), θ_0 là Hệ số ảnh hưởng của khoảng cách từ nơ-ron đến BMU được tính theo công thức:

$$\theta_0 = \exp \frac{-distance^2}{2\sigma_0^2} \quad (07)$$

Về sau, khi mạng đã có trật tự, việc cập nhật quá nhiều nơ-ron sẽ làm mất tính phân biệt chi tiết. Do đó, vùng lân cận cần được thu hẹp, nhằm cho phép SOM điều chỉnh trọng số chính xác tại từng vị trí.

Sự thu hẹp của bán kính vùng lân cận theo số lần huấn luyện được tính theo công thức:

$$\sigma_t = \sigma_0 \exp \left(-\frac{t}{\lambda} \right) \quad (08)$$

Điều chỉnh trọng số W sao cho tiến gần hơn đến V_0 với mức thay đổi được kiểm soát bởi tốc độ học L_0 ,

giá trị L_0 ban đầu thông thường được sử dụng từ (0.1-0.5). Sau khi hiệu chỉnh trọng số BMU, tiếp tục lặp lại quy trình cập nhật trọng số từ công thức (1), với $t=1$ và tăng tiếp tục cho những lần lặp sau. Tốc độ học sẽ giảm dần theo công thức:

$$L_t = L_0 \exp\left(-\frac{t}{\lambda}\right) \quad (09)$$

Trong đó:

- t : là số lần lặp hiện tại
- L_0 là tốc độ học ban đầu
- N : Tổng số lần lặp (number of iterations)
- λ : là hằng số thời gian (constant), được tính bằng:

$$\lambda = \frac{N}{\log \sigma_0} \quad (10)$$

Ban đầu, L_t lớn làm cho trọng số thay đổi nhiều, giúp SOM học nhanh. Về sau, L_t nhỏ dần giúp SOM dần ổn định, tránh bị dao động quá nhiều. Sau khi hiệu chỉnh bán kính theo từng lần lặp tiến hành cập nhật hệ số ảnh hưởng khoảng cách

$$\theta_t = \exp\left(-\frac{\text{distance}^2}{2\sigma_t^2}\right) \quad (11)$$

Cuối cùng là hiệu chỉnh trọng số sao cho gần giống với dữ liệu đầu vào:

$$W(t+1) = W_t + \theta_t L_t (V_t - W_t) \quad (12)$$

Tùy vào dạng lưới mà chọn tổng số lần lặp N phù hợp để tiết kiệm chi phí tài nguyên tính toán và có kết quả chính xác. Thông thường số lần lặp được chọn tối ưu theo công thức thực nghiệm sẽ gấp 500 lần số nơ-ron đầu vào. Tuy nhiên, ở một tập dữ liệu lớn và đa chiều thì số lần lặp này là quá lớn sẽ khó khăn trong việc tính toán. Nên có thể giảm bớt để thuận tiện cho việc tính toán giá trị mong muốn⁸.

Phân cụm phân cấp (Hierarchical clustering algorithm)

Phân cụm phân cấp là một phương pháp chia nhóm dữ liệu mà không cần biết trước có bao nhiêu nhóm mà tùy thuộc vào đặc tính tương đồng của dữ liệu. Thuật toán này được thực hiện bằng cách xây dựng một cấu trúc giống cái cây nhiều tầng, mỗi tầng tương ứng với khoảng cách điểm dữ liệu khác nhau từ đó số cụm cũng được phân chia khác nhau.

Trong Hình 1, nếu cắt trong đoạn đầu từ trên xuống thì được 2 cụm là ABC và DEFG. Nếu cắt từ phần gạch nét đứt sẽ có 3 cụm ABC, DE và FG. Còn tùy thuộc

vào mục đích đầu ra, nếu muốn phân cụm chi tiết hay đơn giản.

Trong những bài toán tính toán phân cụm phân cấp, cách tính khoảng cách trong không gian đa chiều được phân thành 2 loại:

Liên kết đơn (single link): Đây là phương pháp phân cụm dựa trên một cặp điểm gần nhau nhất giữa các cụm, từ đó gộp cụm lại thành một lớn.

Liên kết đơn chính là chọn ra cặp điểm gần nhất của hai cụm A và B so với các điểm còn lại của các cụm khác, từ đó gộp thành cụm AB.

Liên kết toàn bộ (Complete Link): Ngược lại với liên kết đơn, phương pháp này là dựa trên một cặp điểm xa nhất giữa các cụm. Bằng cách chọn cặp điểm nào xa nhất giữa các cụm là gần nhau hơn thì nhóm các cụm đó lại với nhau¹⁰.

DỮ LIỆU

Với dữ liệu đầu vào là các dữ liệu liên quan tới địa vật lý giếng khoan. Hình 3 biểu diễn các giá trị với độ sâu bắt đầu là 500m và độ sâu cuối cùng là 4950m. Bốn dữ liệu là kết quả quá trình ghi lại nhật ký bao gồm là nhật ký gamma ray (GR), nhật ký mật độ (RHOB), nhật ký âm thanh (DT), nhật ký neutron (NPHI).

Lưu ý: Nhật ký âm thanh (DT) bao gồm thời gian truyền sóng nén (DTC) và thời gian truyền sóng cắt (DTS). Trong quá trình nghiên cứu, tác giả thấy trong quá trình phân cụm sơ cấp thì thời gian truyền sóng nén (DTC) sẽ được ưu tiên hơn sóng cắt (DTS). Bởi sóng nén nhảy với độ rộng và các thành phần ma trận đá, điều đó sẽ giúp dễ dàng phân biệt sự khác nhau rõ ràng về loại thạch học.

KẾT QUẢ VÀ THẢO LUẬN

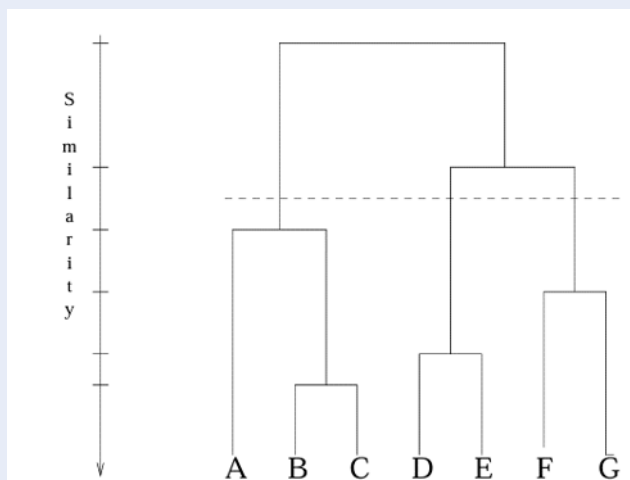
Trong nghiên cứu này, Tác giả đã lựa chọn mạng SOM và thực hiện trên cấu trúc hình lục giác (Hexagonal) và hình cầu (Spherical). Các tham số đầu vào quan trọng cần lưu ý bao gồm là:

- Kích thước của hình lục giác ($n*n$) = $30*30$
- Bản đồ số nút của hình cầu = 642 nút
- Số lần lặp $N=70000$
- Tốc độ học $L_0 = 0.5$

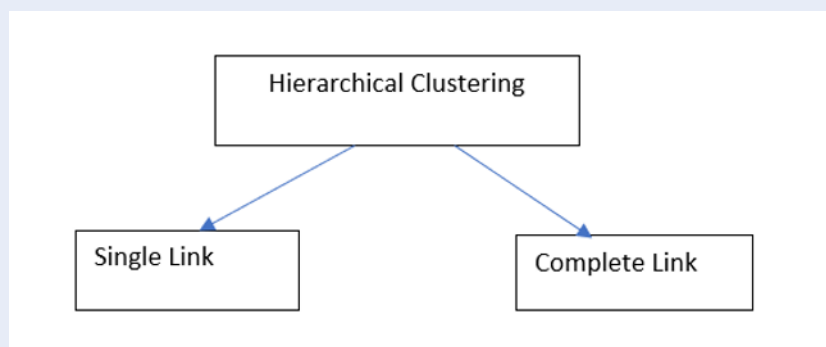
Kết quả của quá trình huấn luyện SOM được biểu diễn ở Hình 4.

Tác giả đã nhận thấy từ bản đồ (Hình 4a và Hình 4b) là những viên đậm thể hiện khoảng cách lớn trong không gian đa chiều giữa hai nút liên kế. Còn những viên nhạt là thể hiện khoảng cách nhỏ và thể hiện 2 nút liên kế có thuộc tính giống nhau.

Ở đây, dựa vào Hình 5 nhận thấy tại phân cụm thứ 6-7 cho thấy giá trị Randomness Ratio dần phẳng hơn, ít



Hình 1: Sơ đồ cây (dendrogram)⁹.



Hình 2: Cách tính toán khoảng cách dữ liệu¹⁰.

sự giảm mạnh cho thấy tại số phân cụm này dữ liệu đã có sự liên kết cấu trúc chặt chẽ. Nên tác giả đã chọn 7 phân cụm, để thực hiện bước phân cụm phân cấp cho 2 cấu trúc.

Dựa trên kết quả phân cụm SOM (Hình 6a và Hình 6b) kết hợp với dữ liệu log (Hình 3), các nhóm thạch học được dự đoán như sau:

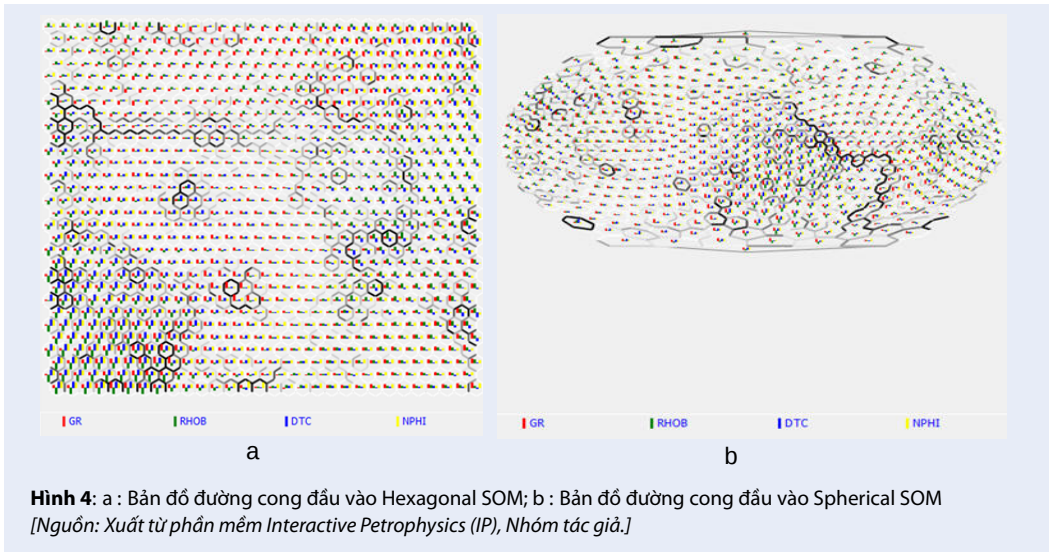
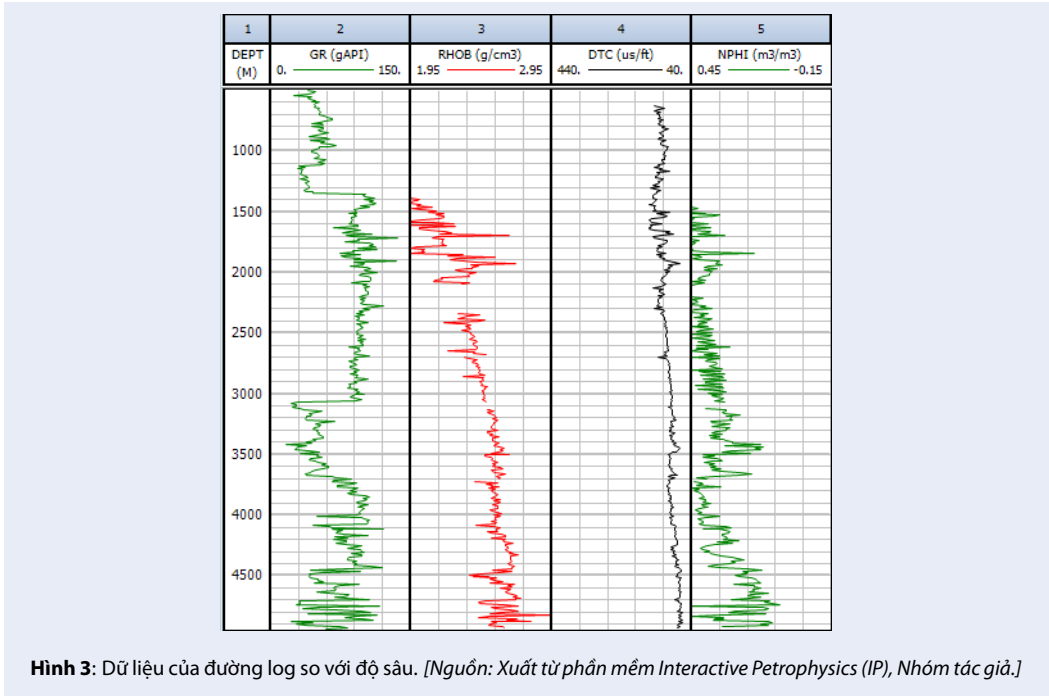
- Cụm 1: Có giá trị GR thấp (< 50 API), DTC trong khoảng $75-80 \mu\text{s}/\text{ft}$, NPHI xấp xỉ 15% và đường RHOB không chênh lệch nhiều. Nên cụm 1 có thể dự đoán là tiềm năng nhất chứa hydrocarbon, loại đá ở đây đa phần sẽ là cát kết sạch.
- Cụm 2: Với tập dữ liệu này là những dữ liệu bị nhiễu hoặc không có giá trị đo được, biểu diễn là -999. Nên cụm 2 sẽ không dự đoán được cụm thạch học vì chưa đủ số liệu.
- Cụm 4, 5, 6 và 7: Đặc trưng bởi giá trị GR cao ($86-134$ API), NPHI ở mức 30–40% và DTC $>$

$110 \mu\text{s}/\text{ft}$. Dự đoán cao thành phần thạch học ở khu vực cụm này đan xen kẽ nhau này sẽ là sét ngậm nước vì giá trị NPHI và GR đều cao ở khu vực các cụm này. Đó cũng chính là đặc trưng của tầng sét. Và đây cũng chính là thành phần chiếm phân cụm lớn nhất của giếng.

- Cụm 3: Có GR từ $55-63$ API, NPHI dao động 10–20%, RHOB khoảng $2,6 \text{ g}/\text{cm}^3$ và DTC từ $69-77 \mu\text{s}/\text{ft}$. Dự đoán tại cụm 3 có giá trị NPHI và RHOB nằm trong khoảng xuất hiện của đá cacbonat, nhưng với DTC nằm trong khoảng giá trị này có thể lẫn ít sét. Và đá cacbonat đa phần có thể dự đoán là dolomite lẫn ít sét.

Để đánh giá độ chính xác của 1 mô hình, tác giả đo độ chính xác bằng tỷ lệ phân loại trung bình của k gần nhất (k – nearest nodes accuracy). Trong đó, giá trị k thay đổi từ $k=1$ đến $k=6$.

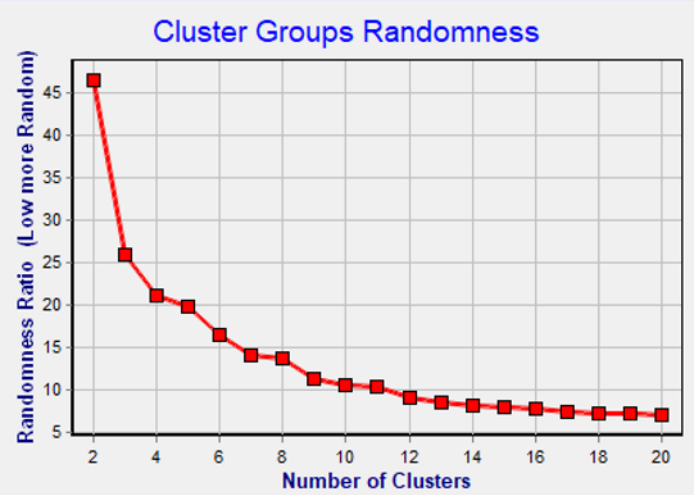
Bảng 1 trình bày kết quả độ chính xác của 2 cấu trúc của SOM. Kết quả cho thấy cả 2 đều đạt kết quả cao



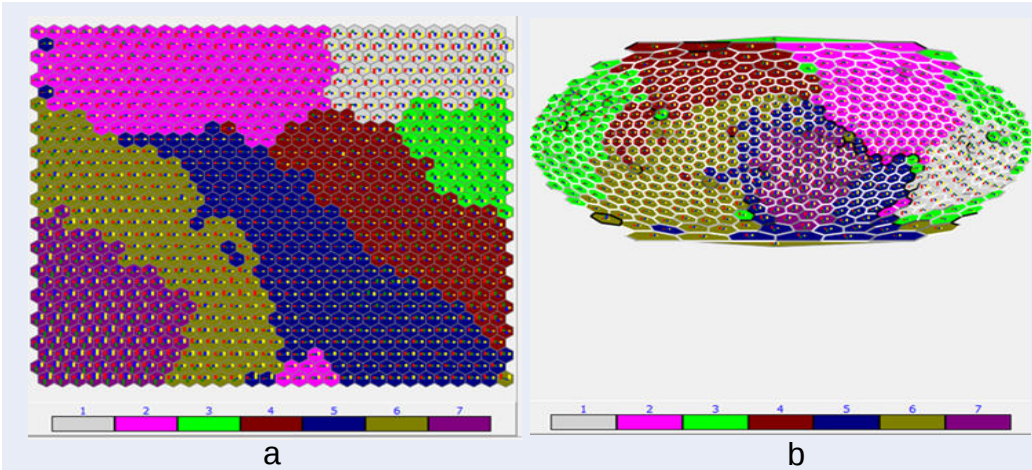
Bảng 1: Độ chính xác của các mô hình SOM hình lục giác và hình cầu dựa trên nút gần nhất và trung bình các nút lân cận.

Gía trị k Loại SOM	Closest Node (k=1)	Average closest 2 node (k=2)	Average closest 3 node (k=3)	Average closest 4 node (k=4)	Average closest 5 node (k=5)	Average closest 6 node (k=6)
Hexagonal SOMs	97.37%	96.59%	96.33%	96.08%	95.87%	95.68%
Spherical SOMs	98.54%	97.42%	96.99%	96.72%	96.62%	96.51%

[Nguồn: Xuất từ phần mềm Interactive Petrophysics (IP), Nhóm tác giả.]



Hình 5: Biểu đồ phân bố cụm theo chỉ số ngẫu nhiên. [Nguồn: Xuất từ phần mềm Interactive Petrophysics (IP), Nhóm tác giả.]



Hình 6: a : Bản đồ phân cụm phân cấp của Hexagonal SOM; b : Bản đồ phân cụm của Spherical SOM. [Nguồn: Xuất từ phần mềm Interactive Petrophysics (IP), Nhóm tác giả.]

Facies	Value	Text Value	Color
1	1	cát kết	
2	2	không xác định	
3	3	dolomite	
4	4	sét	
5	5	sét	
6	6	sét	
7	7	sét	

Hình 7: Kết quả dự đoán thạch học với số màu hiển thị trên bản đồ SOM. [Nguồn: Xuất từ phần mềm Interactive Petrophysics (IP), Nhóm tác giả.]

ở nút gần nhất (Closest Node) với Hexagonal SOM là 97,37% và Spherical SOMs là 98,54%. Khi mở rộng phạm vi đánh giá đối với trung bình các nút lân cận, độ chính xác của cả 2 có xu hướng giảm nhẹ.

Sự giảm nhẹ độ chính xác khi tăng số lượng nút lân cận có thể được giải thích bởi hiệu ứng biên. Trong các bản đồ SOM phẳng, các nút ở rìa có ít láng giềng hơn, dẫn đến sai lệch nhỏ trong quá trình cập nhật trọng số. Điều này không xuất hiện trong Spherical SOM do cấu trúc không biên của bề mặt cầu. Nhìn chung, kết quả của Spherical SOM sẽ nhìn hơn so với Hexagonal SOM ở hầu hết các mức trung bình. Điều này cho thấy cấu trúc của SOM không ảnh hưởng lớn đến quá trình phân cụm thạch học.

KẾT LUẬN

Kết quả nghiên cứu cho thấy thuật toán Self-Organizing Map (SOM) là một công cụ hiệu quả trong việc phân tích và phân cụm dữ liệu địa vật lý giếng khoan. Với khả năng tự tổ chức và bảo toàn cấu trúc lân cận của dữ liệu, SOM giúp xác định rõ ràng các nhóm thạch học có đặc trưng khác nhau, qua đó hỗ trợ đánh giá tiềm năng tầng chứa một cách khách quan và trực quan hơn. So sánh giữa hai cấu trúc bản đồ cho thấy SOM hình cầu (Spherical SOM) cho độ chính xác cao hơn so với SOM hình lục giác (Hexagonal SOM), do hạn chế được ảnh hưởng của hiệu ứng biên trong quá trình huấn luyện.

Tổng thể, nghiên cứu này đã chứng minh tiềm năng ứng dụng rộng rãi của SOM và khẳng định vai trò quan trọng trong việc xử lý và phân tích khối lượng lớn dữ liệu địa vật lý trong ngành dầu khí.

DANH MỤC TỪ VIẾT TẮT

- SOM: Self-Organizing Map (Bản đồ tự tổ chức)
- ANN: Artificial Neural Network (Mạng nơ-ron nhân tạo)
- BMU : Best Matching Unit
- GR: Gamma Ray
- RHOB: Bulk Density
- NPHI: Neutron Porosity
- DT: Sonic Transit Time
- DTC: Compressional Sonic Transit Time (Thời gian truyền sóng dọc)
- DTS: Shear Sonic Transit Time (Thời gian truyền sóng ngang)

ĐÓNG GÓP CỦA CÁC TÁC GIẢ

Trần Viết Việt Anh thực hiện thu thập và xử lý dữ liệu, xây dựng mô hình Self-Organizing Map, phân tích kết quả và soạn thảo bản thảo bài báo. Phạm Sơn Tùng định hướng nghiên cứu, góp ý hoàn thiện phương pháp, kiểm tra tính khoa học của kết quả và chỉnh sửa nội dung bản thảo. Tất cả các tác giả đã đọc và thống nhất với phiên bản cuối cùng của bài báo.

TUYÊN BỐ XUNG ĐỘT LỢI ÍCH

Tác giả cam kết rằng không tồn tại bất kỳ xung đột lợi ích nào liên quan đến việc thực hiện nghiên cứu và công bố kết quả của bài báo này.

LỜI CẢM ƠN

Nghiên cứu được tài trợ bởi Đại học Quốc gia Thành phố Hồ Chí Minh (ĐHQG-HCM) trong khuôn khổ Đề tài mã số C2026-20-20.

TÀI LIỆU THAM KHẢO

1. Kohonen T. The self-organizing map. Proceedings of the IEEE [Internet]. 1990;78(9):1464–80. Available from: <https://doi.org/10.1109/5.58325>; <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=58325>.
2. Phạm TS, Trương HM, Phạm TB. Application of self organizing map in construction, geology and petroleum industry. Science and Technology Development Journal. 2017;20:30–8. Available from: <https://doi.org/10.32508/stdj.v20iK4.1110>.
3. PotoP, Berlec T, Starbek M, Govekar E. SOM-based clustering and optimization of production. Berlin, Heidelberg: Springer; 2011. Available from: https://doi.org/10.1007/978-3-642-23957-1_3.
4. Sarker IH. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science. 2021;2(3):160. Available from: <https://doi.org/10.1007/s42979-021-00592-x>.
5. Ezugwu AE, Ikotun AM, Oyelade OO, Abualigah L, Agushaka JO, Eke CI, et al. A comprehensive survey of clustering algorithms: state-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. Engineering Applications of Artificial Intelligence. 2022;110. Available from: <https://doi.org/10.1016/j.engappai.2022.104743>.
6. Shalev-Shwartz S, Ben-David S. Understanding Machine Learning; 2009. Available from: <https://www.cambridge.org/core/books/understanding-machine-learning/3059695661405D25673058E43C8BE2A6doi:10.1017/CBO9781107298019>.
7. Wu Y, Takatsuka M. Spherical self-organizing map using efficient indexed geodesic data structure. Neural Networks : The Official Journal of the International Neural Network Society. 2006;19(6-7):900–10. Available from: <https://doi.org/10.1016/j.neunet.2006.05.021>.
8. Kohonen T. Essentials of the self-organizing map. Neural Networks : The Official Journal of the International Neural Network Society. 2013;37:52–65. Available from: <https://doi.org/10.1016/j.neunet.2012.09.018>.
9. Xu R, Wunsch D. Survey of clustering algorithms. IEEE Transactions on Neural Networks. 2005;16(3):645–78. Available from: <https://doi.org/10.1109/TNN.2005.845141>.
10. Ackerman M, Ben-David S. A characterization of linkage-based hierarchical clustering. Journal of Machine Learning Research. 2016;17(231):1–17. Available from: <https://doi.org/10.5555/2946645.3053512>.

Applying the Self-Organizing Map algorithm for lithofacies clustering

Pham Son Tung^{1,2,*}, Tran Viet Viet Anh^{1,2}



Use your smartphone to scan this QR code and download this article

¹Department of Drilling and Production Engineering, Faculty of Geology and Petroleum Engineering, Ho Chi Minh University of Technology (HCMUT), 268 Ly Thuong Kiet Street, Dien Hong Ward, Ho Chi Minh City, Vietnam

²Vietnam National University Ho Chi Minh City, Linh Trung Ward, Ho Chi Minh City, Vietnam

Correspondence

Pham Son Tung, Department of Drilling and Production Engineering, Faculty of Geology and Petroleum Engineering, Ho Chi Minh University of Technology (HCMUT), 268 Ly Thuong Kiet Street, Dien Hong Ward, Ho Chi Minh City, Vietnam

Vietnam National University Ho Chi Minh City, Linh Trung Ward, Ho Chi Minh City, Vietnam

Email: phamsontung@hcmut.edu.vn

History

- Received: 17-11-2025
- Revised: 30-01-2026
- Accepted: 03-06-2026
- Published Online: 10-09-2026

DOI : <https://doi.org/10.32508/vnuhcmj-et.v9i3.1557>



Copyright

© VNUHCM Journal. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

ABSTRACT

In the context of the petroleum industry increasingly emphasizing the optimization of geological data analysis and interpretation to enhance exploration and production efficiency, machine learning techniques have demonstrated significant potential in processing and analyzing large volumes of complex, high-dimensional geophysical data. The advancement of unsupervised learning approaches has introduced new opportunities for extracting hidden knowledge from well log datasets without relying solely on subjective expert interpretation. Among these approaches, the Self-Organizing Map (SOM) algorithm has emerged as an effective tool due to its capability to perform clustering while preserving the nonlinear relationships among input variables and visualizing multidimensional data in an interpretable two-dimensional representation.

SOM has been widely applied in various petroleum engineering applications, including lithology classification, sedimentary environment identification, hydraulic flow unit delineation, reservoir property prediction, well log interpretation, and reservoir modeling. By projecting complex data structures onto a topological map, this algorithm enables geoscientists and reservoir engineers to identify characteristic patterns, reveal relationships among geophysical parameters, and support decision-making processes under conditions of data complexity and uncertainty.

This paper presents the application of the Self-Organizing Map algorithm for clustering well log data to identify potential lithological groups. The input dataset consists of selected conventional well logs, which were preprocessed and normalized prior to SOM training and analysis. The results demonstrate that SOM effectively separates data into clusters with similar characteristics, thereby capturing reservoir heterogeneity and facilitating the identification of intervals with potential hydrocarbon-bearing properties. The findings confirm the feasibility and effectiveness of SOM as a tool for geophysical data analysis and provide additional scientific support for well potential assessment and prediction, ultimately improving the reliability of decision-making during petroleum exploration and development activities.

Key words: Self-Organizing Map (SOM), lithological clustering, machine learning, artificial neural networks, well log analysis

Cite this article : Tung P S, Anh T V V. Applying the Self-Organizing Map algorithm for lithofacies clustering. *VNUHCM J. Eng. Technol.* 2026; 9(3): 3294-3301.