TẠP CHÍ
KHÍ TƯỢNG THỦY VĂN

Bài báo khoa học

Dự báo khai thác dầu khí sử dụng thuật giải di truyền (GA) dựa trên việc huấn luyện mạng nơ-ron hồi quy có bộ nhớ ngắn hạn định hướng dài hạn (LSTM)

Phùng Đại Khánh[1], Nguyễn Xuân Huy[1,*]

[1] Khoa Kỹ thuật Địa chất và Dầu Khí, Trường Đại học Bách Khoa, Đại học Quốc Gia Tp.HCM; phungdaikhanh@hcmut.edu.vn; nxhuy@hcmut.edu.vn

*Tác giả liên hệ: nxhuy@hcmut.edu.vn; Tel.: +84–909453698

Ban Biên tập nhận bài: 8/12/2021; Ngày phản biện xong: 10/3/2022; Ngày đăng bài: 25/4/2022

Tóm tắt: Một trong những nhiệm vụ then chốt của việc quản lý khai thác mỏ dầu khí là sử dụng dữ liệu lịch sử khai thác để dự báo sản lượng khai thác trong tương lai và đánh giá trữ lượng trong quá trình lên kế hoạch phát triển mỏ dầu khí. Gần đây, lĩnh vực học máy, học sâu đã giải quyết được những hạn chế của các phương pháp dự báo truyền thống là phức tạp và tốn nhiều thời gian. Với sự gia tăng theo thời gian lượng dữ liệu khai thác, thì cách tiếp cận học sâu này thể hiện tính hợp lý và độ tin cậy cũng như sự chính xác cao trong công tác dự báo khai thác. Bài báo này đề xuất một cách tiếp cận học sâu để dự báo lượng khai thác dầu khí bằng các mạng nơ-ron hồi quy có bộ nhớ ngắn hạn định hướng dài hạn sâu (Deep LSTMs - DLSTM). Thuật giải di truyền (Genetic Algorithm-GA) được kết hợp sử dụng để tối ưu hóa mạng DLSTM. Cách tiếp cận này được vận dụng để dự báo khai thác cho mỏ STD, bể Cửu Long. Kết quả dự báo chính xác đã thể hiện được hiệu quả và sự đúng đắn của cách tiếp cận và phương pháp dự báo. Cách tiếp cận này có thể được áp dụng cho các mỏ tương tự trong khu vực.

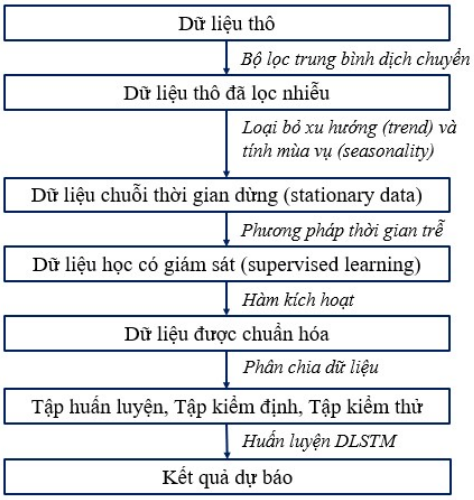
Từ khóa: Học máy; Học sâu; Thuật giải di truyền; Bộ nhớ ngắn hạn định hướng dài hạn sâu; Dầu khí.

1. Mở đầu

Dữ liệu chuỗi thời gian trong thực tế rất phức tạp, có nhiều nhiễu nên rất khó mô tả, chẳng hạn như các phương trình phân tích dựa trên các tham số của đường cong suy giảm áp suất (*Decline Curve Analysis–DCA*). Nhược điểm chính của các phương pháp phân tích truyền thống là chủ yếu dựa trên loại dữ liệu chủ quan. Nói cách khác, việc lựa chọn độ dốc thích hợp và sau đó điều chỉnh các tham số của mô hình mô phỏng số, giữ lại các giá trị hợp lý, và cuối cùng, đưa ra các diễn giải về động thái thay đổi sản lượng khai thác của mỏ dầu khí. Nhưng, các đặc điểm về địa chất và đặc tính chất lưu trong tầng chứa thường là các ứng xử phi tuyến và bất đồng nhất về bản chất, do đó dữ liệu chuỗi thời gian sẽ thể hiện quá trình bộ nhớ dài. Bên cạnh đó, việc xây dựng mô hình vỉa bằng phương pháp số truyền thống cho một mỏ cụ thể với hàng trăm giếng khoan dựa trên các nguyên tắc vật lí vỉa và địa chất là một quá trình tốn kém và tốn thời gian. Vì vậy, cách tiếp cận sử dụng mạng các Bộ nhớ ngắn hạn định hướng dài hạn sâu (Deep Long Short-Term Memory Networks - DLSTM) sẽ là một giải pháp hợp lý trong trường hợp này để dự báo sản lượng dầu khí khai thác. Để tăng hiệu quả của mô hình DLTSM, nghiên cứu này cũng đã sử dụng thuật giải di truyền (*Genetic Algorithm–GA*) trong quá trình tính toán xác định các siêu thông số (*hyper-parameters*) cho mạng các DLSTM. Cách tiếp cận kết hợp GA và DLSTM này đã đưa ra giải pháp nhanh chóng và chính xác nhất để dự báo sản lượng khai thác dầu khí.

2. Chuẩn bị dữ liệu và phương pháp nghiên cứu

Các bước thực hiện công việc huấn luyện DLSTM được thể hiện ở Hình 1.



Hình 1. Quy trình thực hiện huấn luyện DLSTM.

2.1. Bộ dữ liệu

Dữ liệu được sử dụng trong nghiên cứu này bao gồm các dữ liệu thô thu được từ *Lô X*, mỏ *STD*, bể *Cửu Long*. Các dữ liệu đầu vào liên quan đến quá trình khai thác bao gồm áp suất (*pressure*), nhiệt độ (*temperature*), lưu lượng (*flow rate*) và van điều tiết (*top-side choke valve)* mở trong khoảng thời gian 10 phút.

Tiền xử lý dữ liệu

Dữ liệu được sử dụng là dữ liệu khai thác thô của mỏ *STD*, vì vậy rất có thể bao gồm các yếu tố ảnh hưởng của nhiễu. Như vậy, việc sử dụng dữ liệu thô trong quá trình học của mạng nơ-ron là không phù hợp vì sẽ dẫn đến tốc độ học (*learning rate*) của mạng rất thấp. Do đó, một công tác tiền xử lý dữ liệu bao gồm bốn bước đã được sử dụng trước khi dữ liệu khai thác thô được đưa vào luyện mạng trong nghiên cứu này. Các bước tiền xử lý bao gồm 4 bước được mô tả bên dưới:

Bước 1: Loại bỏ nhiễu từ dữ liệu thô

Sử dụng bộ lọc trung bình dịch chuyển (*moving average filter*) để làm mịn dữ liệu thô và loại bỏ nhiễu, theo cách tương tự như được mô tả trong [1]. Cụ thể, bộ lọc này cung cấp trung bình trọng số của các điểm dữ liệu trong quá khứ cho bộ dữ liệu khai thác theo thời gian trong một chu kì năm điểm để tạo ra ước lượng làm mịn dữ liệu của một chuỗi thời gian. Phải nhất thiết kết hợp bước này để giảm nhiễu ngẫu nhiên trong bộ dữ liệu bằng cách giữ lại phản hồi tốt nhất liên quan đến dữ liệu thô.

Bước 2: Chuyển đổi dữ liệu thô thành dữ liệu chuỗi thời gian dừng (Stationary data)

Dữ liệu chuỗi thời gian thường thể hiện là chuỗi dữ liệu không cố định, trên thực tế, sẽ thể hiện một xu hướng cụ thể [2]. Tất nhiên, dữ liệu chuỗi thời gian dừng sẽ dễ dàng hơn để mô hình hóa và có thể dẫn đến dự báo khéo léo hơn. Dữ liệu chuỗi thời gian là dừng nếu chúng không có thêm xu hướng (*trend)* và tính mùa vụ (seasonal). Các đặc tính thống kê trên chuỗi thời gian là nhất quán theo thời gian, ví dụ như giá trị trung bình (mean) và phương sai (variance). Khi dữ liệu chuỗi thời gian ở trạng thái dừng thì chúng có thể dễ dàng mô hình hóa với độ chính xác cao hơn. Chuỗi thời gian dừng (gọi tắt là chuỗi dừng) sẽ không bao hàm các yếu tố xu thế.

Sau khi bước tiền xử lý, thuộc tính xu hướng trong dữ liệu được loại bỏ; cho dù đó là xu hướng tăng hay giảm. Sau đó, xu hướng được thêm trở lại mô hình dự báo để trả lại kết quả dự báo theo tỷ lệ *(scale*) ban đầu và tính toán một giá trị sai số tương đương. Một cách tiêu

chuẩn để loại bỏ xu hướng do sự khác biệt của dữ liệu. Đó là do có sự quan sát từ bước thời gian hiện tại (t) trừ cho bước thời gian trước đó (t-1) [2].

Bước 3: Chuyển đổi dữ liệu thành học tập có giám sát (Supervised learning)

Đầu tiên sử dụng dự báo trước một bước, trong đó bước tiếp theo (t + 1) được dự đoán. Dữ liệu chuỗi thời gian được chia thành đầu vào (x) và đầu ra (y) bằng phương pháp độ trễ thời gian (*lag time method*), và cụ thể, trong nghiên cứu này sử dụng các kích thước khác nhau của độ trễ, từ lag1 đến lag6.

Bước 4: Chuẩn hóa dữ liệu

Giống như các mạng nơ-ron khác, DLSTM sử dụng hàm kích hoạt (*activation function*) với mong đợi dữ liệu sẽ nằm trong phạm vi tỷ lệ của hàm. Các hàm kích hoạt mặc định cho LSTM là tiếp tuyến hyperbolic (tanh), trong đó các giá trị đầu ra của nó nằm giữa −1 và 1. Đây là phạm vi ưa thích cho dữ liệu chuỗi thời gian. Dữ liệu được chuyển đổi thu nhỏ lại để trả về giá trị dự báo về tỷ lệ (*scale*) ban đầu.

Phân chia dữ liệu

Bộ dữ liệu (chứa 227 kết quả từ dữ liệu sản lượng dầu khai thác) được chia thành 2 tập, tập thứ nhất (80% dữ liệu, chứa 182 kết quả) được sử dụng để xây dựng, huấn luyện, xác thực mô hình dự báo và tập thứ hai (20% dữ liệu còn lại, chứa 45 kết quả) được sử dụng làm tập kiểm tra hiệu quả của mô hình dự báo.

Thực hiện các kịch bản

Việc thực hiện các mô hình đề xuất DLSTM bao gồm hai kịch bản khác nhau, bao gồm: kịch bản tĩnh và kịch bản động. Trong mô hình kịch bản tĩnh, mô hình dự báo được làm phù hợp với dữ liệu huấn luyện và dự báo mỗi bước thời gian mới một lần tại một thời điểm với dữ liệu kiểm tra (*testing data*). Trong mô hình kịch bản động, mô hình dự báo được cập nhật mỗi bước thời gian khi có thêm các quan sát mới từ các tập dữ liệu kiểm tra (*testing data*). Hay nói cách khác, kịch bản động sử dụng giá trị dự đoán trước đó của biến độc lập để tính toán cho lần kế tiếp, trong khi dự báo tĩnh sử dụng giá trị thực cho mỗi lần dự báo tiếp theo.

Phương pháp huấn luyện mạng DLSTM

Trong việc huấn luyện mạng DLSTM, nghiên cứu này đưa vào sử dụng thuật toán di truyền (*Genetic Algorithm–GA*) để đưa ra lựa chọn tối ưu cho siêu tham số (*hyper-parameter*) của mô hình được đề xuất. Số lượng các siêu tham số có được dựa trên kịch bản được thực hiện. Đối với các kịch bản tĩnh, có 3 loại siêu tham số: số lượng epoch, số lượng lớp ẩn, và lag size. Đối với mô hình động sẽ có 4 siêu tham số: 3 tham số giống của mô hình tĩnh và thêm 1 tham số nữa là số lần cập nhật *(number of updates*), có nghĩa là số lần cập nhật mỗi bước thời gian của mô hình dự báo khi có thêm các quan sát mới từ dữ liệu kiểm tra.

Thuật toán di truyền (GA)

Thuật toán di truyền giúp tìm ra các giải pháp chính xác hoặc gần đúng trong các vấn đề tối ưu hóa, tìm kiếm và học tập. Lấy cảm hứng từ lý thuyết tiến hóa của Darwin, các giải pháp của thuật toán di truyền phát triển theo thời gian bằng các phương pháp chọn lọc (*selection*), lai tạo (*crossover*) và đột biến (*mutation*), tương tự như các hoạt động từ tự nhiên. Mỗi giải pháp vấn đề được mã hóa thành một chuỗi có độ dài hữu hạn (nhiễm sắc thể) trên một bảng chữ cái hữu hạn. Thuật toán xác định một quần thể các giải pháp hoặc các cá thể.

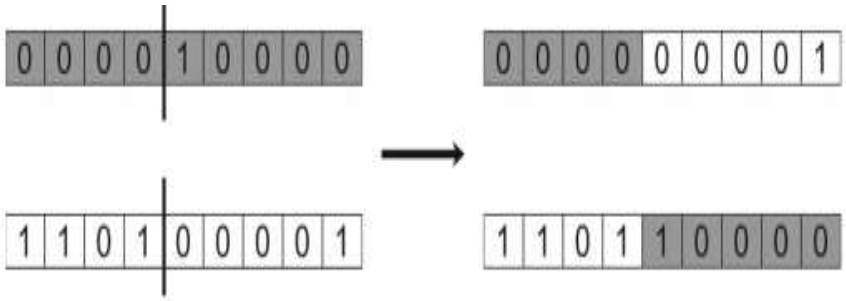
Khác biệt quan trọng giữa tìm kiếm của GA và các phương pháp tìm kiếm khác là thuật giải duy truyền duy trì và xử lý một tập các lời giải, gọi là một quần thể (*population*). Trong GA, việc tìm kiếm giả thuyết thích hợp được bắt đầu với một quần thể, hay một tập hợp có chọn lọc ban đầu của các giả thuyết. Các cá thể của quần thể hiện tại khởi nguồn cho quần thể thế hệ kế tiếp bằng các hoạt động lai ghép và đột biến ngẫu nhiên–được lấy mẫu sau các quá trình tiến hóa sinh học. Ở mỗi bước, các giả thuyết trong quần thể hiện tại được ước lượng liên hệ với đại lượng thích nghi, với các giả thuyết phù hợp nhất được chọn theo xác suất là các hạt giống cho việc sản sinh thế hệ kế tiếp, gọi là cá thể (*individual*). Cá thể nào phát triển hơn, thích ứng hơn với môi trường sẽ tồn tại và ngược lại sẽ bị đào thải. GA có thể dò tìm thế hệ mới có độ thích nghi tốt hơn. GA giải quyết các bài toán quy hoạch toán học thông qua các

quá trình cơ bản: lai tạo (*crossover*), đột biến (*mutation*) và chọn lọc (*selection*) cho các cá thể trong quần thể. Dùng GA đòi hỏi phải xác định được: khởi tạo quần thể ban đầu, hàm đánh giá các lời giải theo mức độ thích nghi–hàm mục tiêu, các toán tử di truyền tạo hàm sinh sản.

Chọn lọc (*Selection*). Tất cả các giả thuyết trong quần thể được đánh giá dựa trên hàm phù hợp (*fitness function*). Quần thể mới được tạo ra bằng việc lựa chọn xác suất các cá thể tốt nhất từ quần thể hiện tại.

Lai tạo (*Crossover*): Thuật toán kết hợp hoặc hợp nhất việc mã hóa của hai tập hợp ban đầu để tạo ra các cá thể mới. Một toán tử chung, được gọi là phép lai tạo một điểm (one-point crossover), chọn ngẫu nhiên một điểm trong chuỗi nhiễm sắc thể và sau đó trao đổi các giá trị từ điểm đó để tạo ra hai tập hợp con (Hình 2).

Đột biến (*Mutation)*. Thuật toán thay đổi xác suất một số yếu tố của chuỗi mã hóa, tìm kiếm các đặc điểm mới không có sẵn ở các cá thể từ các quần thể trước đó. Đột biến ngăn một thuật toán di truyền hội tụ nhanh chóng mà không cần thăm dò đủ trong không gian tìm kiếm.



Hình 2. Sự lai tạo 1 điểm để hợp nhất 2 giải pháp trọng GA [3].

Mô tả chi tiết thuật toán GA:

GA (*Fitness, Fitness_thresshold, p, r, m*)

Fitness: Hàm gán điểm đánh giá độ phù hợp cho mỗi giải thuyết được đưa vào.

Fitness_threshold: Ngưỡng điểm đánh giá phù hợp để dừng thuật toán.

p: Số lượng giả thuyết trong quần thể.

r: Tỉ lệ số giả thuyết được lai tạo trong quần thể.

m: Tỉ lệ đột biến.

Quần thể ban đầu: P ← Khởi tạo p giả thuyết ngẫu nhiên.

Đánh giá: Với mỗi giả thuyết h trong P, tính toán Fitness(h)

Trong khi [maxhFitness(h)]<Fitness_threshold], tiến hành

Khởi tạo một thế hệ mới Ps:

1. Lựa chọn: Lấy xác suất (1−r)p thành viên của P để thêm vào Ps. Xác suất $Pr(h_i)$ của việc chọn giả thuyết $Pr(h_i)$ từ P được tính bởi (1):

$$\Pr\left(h_i\right)=\frac{\text{Fitness}(h_t)}{\sum_p^{j=1}\text{Fitness}(h_t)} \quad (1)$$

2. Lai tạo: Lựa chọn theo xác suất r∗p/2 cặp giả thuyết từ P, dựa vào $Pr(h_i)$ ở phía trên. Với mỗi cặp, (h_l,h_2), tạo ra hai giả thuyết con bằng phép lai tạo. Thêm tất cả các giả thuyết con vào P_s.

3. Đột biến: Lựa chọn m phần trăm thành viên từ P_s, với xác suất như nhau. Với mỗi thành viên được chọn, tiến hành đảo ngược một bit bất kì trên tập bit biểu diễn giả thuyết đó.

4. Cập nhật. P←P_s

5. Đánh giá: với mỗi h trong P, tính toán Fitness (h).

• Trả về giả thuyết thuộc P có độ phù hợp cao nhất.

3. Kết quả nghiên cứu và thảo luận

Kết quả tốt nhất của các kịch bản tĩnh và kịch bản động của mô hình DLSTM được thể hiện trong Bảng 1. Mỗi bảng cho biết giá trị siêu tham số (*hyper-parameter*) tối ưu được lựa chọn theo thuật toán di truyền (GA) như miêu tả phần 2.1.4. Cả 2 bảng đều cho kết quả tốt nhất của DLSTM với sử dụng 3 lớp LSTM cho kịch bản tĩnh và 2 lớp LSTM cho kịch bản động.

Khi sử dụng mô hình DLSTM với kịch bản tĩnh (Bảng 1), sai số bình phương trung bình căn thức (*Root Mean Square Error–RMSE*) và sai số phần trăm bình phương trung bình căn thức (*Root Mean Square Percentage Error*) lần lượt là 0,209 và 2,995, sai số này là thấp nhất ứng với trường hợp số lớp LSTM được thực hiện là 3 lớp.

Khi sử dụng mô hình DLSTM với kịch bản động (Bảng 2), sai số RMSE = 0,219 và sai số RMSPE = 3,124 là thấp nhất, ứng với trường hợp số lớp LSTM được xếp chồng là 2 lớp. Khi tiếp tục tăng số lớp LSTM này lên 3, thì sai số cũng tăng lên với RMSE 0,257, RMSPE = 3,124.

Có thể thấy, khi thực hiện kịch bản động, bằng việc thêm vào các cập nhật khi có quan sát mới, cho được kết quả tối ưu hơn so với mô hình DLSTM kịch bản tĩnh, vì chỉ dừng lại ở 2 lớp LSTM, tiết kiệm thời gian và chi phí trong trường hợp dữ liệu lớn.

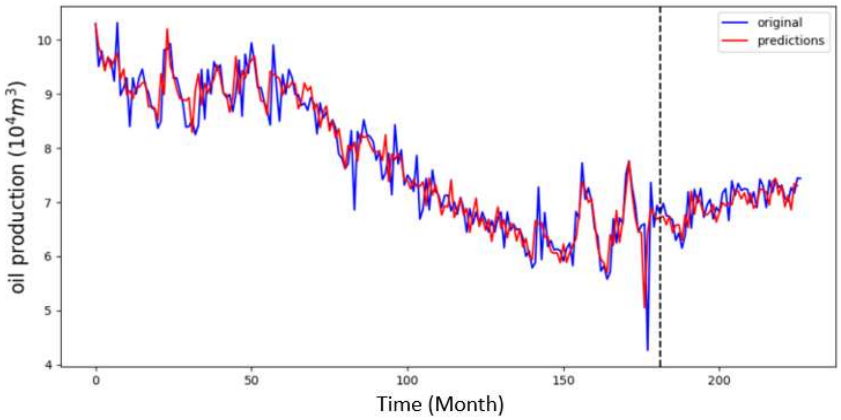
Bảng 1. Kết quả tốt nhất của mô hình DLSTM với kịch bản tĩnh.

No. of layer	No. of hidden units	No. of Epochs	Lag	RMSE	RMSPE
1	[4]	953	5	0,234	3,337
2	[4, 2]	787	5	0,227	3,253
3	[5, 4, 2]	800	5	0,209	2,995

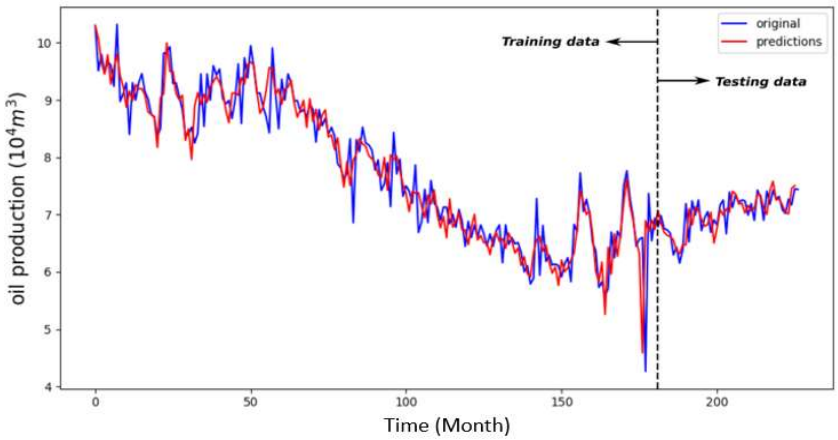
Bảng 2. Kết quả tốt nhất của mô hình DLSTM với kịch bản động.

No. of layer	No. of hidden units	No. of Epochs	Lag	Update	RMSE	RMSPE
1	[3]	1352	3	1	0,267	3,783
2	[4, 5]	1187	5	1	0,219	3,124
3	[4, 3, 3]	403	5	2	0,257	3,637

Mối quan hệ giữa dữ liệu khai thác và dữ liệu dự báo được thể hiện trong Hình 3 và Hình 4. Các kết quả dự báo (đường màu đỏ) hầu như trùng khớp với dữ liệu khai thác thực (đường màu xanh dương), từ đó cho thấy, tính hiệu quả của mô hình DLSTM trong việc dự đoán dữ liệu chuỗi thời gian trong khai thác dầu khí là một lời giải tối ưu. Cả 2 kịch bản đều cho sai số có thể chấp nhận được.



Hình 3. Kết quả dự báo so với dữ liệu khai thác sử dụng mô hình DLSTM-tĩnh.



Hình 4. Kết quả dự báo so với dữ liệu khai thác sử đụng mô hình DLSTM-động.

4. Kết luận

Trong nghiên cứu này đã phát triển một mô hình dự báo đầy hứa hẹn có thể được sử dụng trong phần lớn các vấn đề dự báo chuỗi thời gian. Tuy nhiên, chỉ tiến hành thử nghiệm cụ thể trong trường hợp dữ liệu chuỗi thời gian trong lĩnh vực dầu khí. Mô hình đề xuất là kiến trúc sâu của các mạng nơ-ron có bộ nhớ ngắn hạn định hướng dài hạn (*Long -Short Term Memory - LSTM*), được ký hiệu là DLSTM.

Các kết quả cho thấy rằng, cả hai kịch bản DLSTM tĩnh và động, đều cho kết quả dự báo sản lượng dầu khai thác với sai số có thể chấp nhận được. Cho trường hợp DLSTM kịch bản tĩnh, sai số bình phương trung bình căn thức (*Root Mean Square Error–RMSE*) và sai số phần trăm bình phương trung bình căn thức (*Root Mean Square Percentage Error*) lần lượt là 0,209 và 2,995, sai số này là thấp nhất ứng với trường hợp số lớp LSTM được thực hiện là 3 lớp. Cho mô hình DLSTM kịch bản động, sai số RMSE = 0,219 và sai số RMSPE = 3,124 là thấp nhất, ứng với trường hợp số lớp LSTM được xếp chồng là 2 lớp. Riêng cách xây dựng mô hình DLSTM kịch bản động, bằng việc thêm vào các cập nhật khi có quan sát mới, thì mô hình cho được kết quả tối ưu hơn so với mô hình DLSTM kịch bản tĩnh, vì chỉ dừng lại ở 2 lớp LSTM, tiết kiệm thời gian và chi phí trong trường hợp dữ liệu lớn.

Mô hình chứng tỏ rằng, việc xếp chồng (*stacking*) nhiều lớp LSTM hơn đảm bảo phục hồi các hạn chế của kiến trúc mạng nơ-ron nông, đặc biệt, khi bộ dữ liệu chuỗi thời gian dài được sử dụng. Ngoài ra, mô hình sâu được đề xuất có thể mô tả mối quan hệ phi tuyến giữa đầu vào và đầu ra của hệ thống, đặc biệt trong trường hợp dữ liệu chuỗi thời gian trong dầu khí là không đồng nhất, đầy phức tạp và vẫn còn thiếu nhiều dữ liệu.

Dự đoán chính xác và hiệu suất học được hiển thị trong kết quả trên chỉ ra rằng mô hình DLSTM sâu được đề xuất đủ điều kiện để được áp dụng trong các vấn đề dự báo phi tuyến trong ngành dầu khí. Trong các kế hoạch nghiên cứu trong tương lai, nhóm sẽ tiến hành nghiên cứu hiệu suất của DLSTM trong các vấn đề dự báo khác, đặc biệt là trong các trường hợp bao gồm dữ liệu chuỗi thời gian đa biến.

So với cách tiếp cận mô hình mô phỏng vỉa bằng phương pháp số truyền thống, cách tiếp cận sử dụng DLSTM đưa ra giải pháp nhanh hơn, và là một kỹ thuật thay thế để đưa ra dự báo sản lượng nhanh chóng và đáng tin cậy bên cạnh các mô phỏng số và thực nghiệm.

Đóng góp của tác giả: Xây dựng ý tưởng nghiên cứu: N.X.H., P.Đ.K.; Phân tích số liệu: N.X.H., P.Đ.K.; Viết bản thảo bài báo: N.X.H., P.Đ.K.; Chỉnh sửa bài báo: N.X.H.

Lời cảm ơn: Nghiên cứu này được tài trợ bởi Bộ Khoa học và Công nghệ Việt Nam, trong khuôn khổ đề tài mã số NĐT.48.KR/18. Chúng tôi xin cảm ơn Trường Đại học Bách Khoa, ĐHQG-HCM đã hỗ trợ thời gian và phuong tiện vật chất cho nghiên cứu này. Ngoài ra,

chúng tôi rất cảm on sự hỗ trợ từ Tổng Công ty thăm dò và Khai thác Dầu Khí (PVEP) đã tạo điều kiện cung cấp các số liệu cần thiết để hoàn thành bài báo này.

Lời cam đoan: Tập thể tác giả cam đoan bài báo này là công trình nghiên cứu của tập thể tác giả, chưa được công bố ở đâu, không được sao chép từ những nghiên cứu trước đây; không có sự tranh chấp lợi ích trong nhóm tác giả.

Tài liệu tham khảo

1. Chakra, N.C.; Song, K.Y.; Gupta, M.M.; Saraf, D.N. An innovative neural forecast of cumulative oil production from a petroleum reservoir employing higher-order neural networks (HONNs). *J. Pet. Sci. Eng.* **2013**, *106*, 18–33.
2. Cryer, J.D.; Chan, K.S. Time Series Analysis, 2nd ed., Springer Texts in Statistics, Springer, New York, 2008.
3. Frausto-Solís, J.; Chi-Chim, M.; Sheremetov, L. Forecasting Oil Production Time Series with a Population-Based Simulated Annealing Method. *Arab J. Sci. Eng.* **2015**, 40(4), 1081–1096.

Thuật Ngữ và Danh Mục Viết Tắt

Decline Curve Analysis (DCA): Phân tích đường cong suy giảm

Deep Long-Short Term Memory (DLSTM): Bộ nhớ ngắn hạn định hướng dài hạn sâu

Learning rate: Tốc độ học

Moving average filter: Bộ lọc trung bình dịch chuyển

Stationary data: Dữ liệu chuỗi thời gian dừng

Lag time method: Phương pháp thời gian trễ

Activation Function: Hàm kích hoạt

Genetic Algorithm (GA): Thuật toán di truyền

Hyper-parameter: Siêu tham số

A Petroleum Production Forecasting Method Using Long Short-Term Memory Networks (LSTMs) and Genetic Algorithm (GA)

Phung Dai Khanh[1], Nguyen Xuan Huy[1*]

[1] Faculty of Geology and Petroleum Engineering, Ho Chi Minh City University of Technology (HCMUT); phungdaikhanh@hcmut.edu.vn; nxhuy@hcmut.edu.vn

Abstract: One of the key tasks of oil and gas field management is to use historical data to forecast future production and evaluate reserves during oil field development planning. Recently, the area of machine learning, deep learning has addressed the limitations of traditional forecasting methods that are complicated and time-consuming. For the increase over time in the amount of historical production data, the deep learning approach shows the reasonableness and reliability as well as the high accuracy in petroleum production prediction. This paper proposes a deep learning approach to forecast oil and gas production by utilizing Deep Long Short-Term Memory Networks (Deep LSTMs - DLSTMs). Genetic Algorithm (GA) is also used to optimize the DLSTM network. This approach is applied to forecast the amount of production for the STD reservoir, Cuu Long basin. The accurate forecasting results illustrate the effectiveness and accuracy of the forecasting approaching method. This approach can be applied to similar reservoirs in the region.

Keywords: Deep Long Short-Term Memory Networks; Genetic Algorithm; Oil Production; Decline Curve Analysis; Machine Learning.