Kỷ yếu Hội nghị Khoa học Quốc gia lần thứ IX "Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR'9)"; Cần Thơ, ngày 4-5/8/2016
DOI: 10.15625/vap.2016.00015

DEEP LEARNING: ỨNG DỤNG CHO DỰ BÁO LƯU LƯỢNG NƯỚC ĐẾN HỒ CHỨA HÒA BÌNH

Trương Xuân Nam[1], Nguyễn Thanh Tùng[1]
[1] Khoa Công nghệ thông tin, Trường Đại học Thủy lợi
{namtx, tungnt}@tlu.edu.vn

TÓM TẮT—*Việc dự báo lưu lượng nước về hồ Hoà Bình có vai trò rất quan trọng cho công tác vận hành công trình thủy lợi, thủy điện mục với tiêu phòng lũ cho đồng bằng sông Hồng, góp phần phòng tránh hạn hán và ổn định năng lượng điện cho cả nước. Trong bài báo này, chúng tôi ứng dụng phương pháp Deep learning dự báo lượng nước về hồ Hòa Bình trước 10 ngày. Kết quả thực nghiệm cho thấy mô hình dự báo tìm được có chất lượng dự báo vượt trội so với các phương pháp học máy khác như máy véctơ hỗ trợ, rừng ngẫu nhiên, LASSO, cây quyết định, k láng giềng gần nhất; hệ số xác định bội R^2 của phương pháp Deep learning đạt tới 92%. Nghiên cứu này cũng mở ra hướng ứng dụng mới cho các bài toán dự báo chuỗi thời gian khi dùng Deep learning, phương pháp này giúp cải thiện độ chính xác của mô hình và có thể ứng dụng rộng rãi trong các lĩnh vực kinh tế, xã hội tại Việt Nam.*

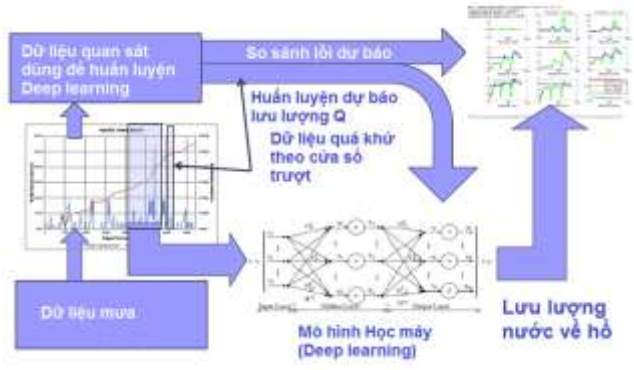
Từ khóa— *Deep learning, mạng nơron, máy véctơ hỗ trợ, rừng ngẫu nhiên, LASSO, khai phá dữ liệu, học máy.*

I. ĐẶT VẤN ĐỀ

Hồ chứa Hòa Bình là hồ chứa lớn nhất Việt Nam nằm trên dòng sông Đà, cách Hà Nội khoảng 70km, hệ thống thủy điện của hồ Hòa Bình được đưa vào vận hành từ năm 1990 giữ vai trò quan trọng trong kiểm soát lũ, hạn hán và cung cấp điện cho đồng bằng châu thổ sông Hồng. Dung tích hồ là $9.6*10^9$ m^3 với khả năng xả lớn nhất đạt 2,400 m^3/s. Điều tiết nước trong mùa lũ luôn là bài toán quan trọng trong vận hành hồ chứa, việc dự báo chính xác lưu lượng nước đổ về hồ chứa trước một khoảng thời gian chủ động luôn là thách thức lớn đối với nhà quản lý trong công tác vận hành tổ hợp công trình tại thủy điện Hòa Bình.

Dự báo chính xác lưu lượng là yêu cầu chủ yếu trong việc xây dựng thành công một hệ thống quản lý và giảm nhẹ ảnh hưởng của lũ, an toàn hồ đập trong một lưu vực sông. Các phương pháp học máy tiếp cận bài toán dự báo lưu lượng nước đổ về hồ chứa đã được nghiên cứu và đạt được nhiều kết quả khả quan [17, 18, 19]. Mạng nơron (neural) nhân tạo (ANN) được chứng minh bằng thực nghiệm cho thấy là một trong các phương pháp hiệu quả nhất và được dùng nhiều trong tính toán dự báo dạng chuỗi thời gian đối với bài toán thuộc dạng này [18, 20-24]. Trong những năm gần đây, phương pháp Deep learning (Học sâu) dựa trên nền tảng mạng nơron nhân tạo đang phát triển rất nhanh và thu hút đông đảo cộng đồng nghiên cứu tham gia. Trong nghiên cứu này, phương pháp Deep learning được nghiên cứu để phân tích, dự báo trung hạn lưu lượng nước đổ về hồ Hòa Bình trước 10 ngày, từ đó trợ giúp công tác hỗ trợ ra quyết định trong điều hành hồ chứa; phương pháp Deep learning cũng có thể mở rộng với các bài toán dự báo lưu lượng tại các hồ chứa khác của Việt Nam phục vụ phòng, tránh và giảm nhẹ thiên tai.

Bài toán dự báo lượng nước đổ về hồ Hòa Bình thường dựa trên vào các dữ liệu quan sát được trong quá khứ và các yếu tố tác động đến lưu lượng nước đổ về hồ. Dự báo lưu lượng nước sẽ chịu tác động của nhiều yếu tố ảnh hưởng đến kết quả dự báo như mưa, dòng chảy, địa hình, thảm phủ thực vật, độ ẩm, khí hậu và các tác động của con người trên lưu vực,… Trong những yếu tố đó thì mưa đóng vai trò quan trọng nhất, còn các yếu tố về địa chất, thổ nhưỡng, thảm phủ thực vật ít thay đổi, nghiên cứu này chưa xét đến yếu tố khí hậu và tác động của con người làm thay đổi lưu vực. Chúng tôi tập trung nghiên cứu vào xây dựng mô hình Deep learning cải thiện chất lượng bài toán dự báo lưu lượng nước đến hồ Hòa Bình dựa trên số liệu mưa quan trắc được và các quan sát lưu lượng nước về hồ trong quá khứ.



Hình 1. Mô hình dự báo sử dụng ANN.

Các số liệu quan trắc về lượng mưa là những yếu tố quan trọng ảnh hưởng trực tiếp đến chất lượng dự báo. Hình 1 mô tả quá trình thu thập số liệu quan trắc theo chuỗi thời gian, thông thường ta xét trong 1 khoảng thời gian cố định (cửa sổ trượt) có dữ liệu quan sát phản ánh đủ những kịch bản dự báo. Những dữ liệu này được gọi chung là tập dữ liệu huấn luyện hay dữ liệu để học mô hình dự báo. Ta ký hiệu tập dữ liệu đầu vào này là $D = \{(X_1, Q_1), (X_2, Q_2), .. (X_N, Q_N)$, trong đó N là số lượng mẫu quan sát được trong quá khứ, tập biến đầu vào X (predictors) là các số liệu quan trắc về lượng mưa và biến đích Q (response feature) lưu giá trị quan sát của lưu lượng nước đổ về hồ Hòa Bình. Xét mô hình dự báo tổng quát để ước lượng lưu lượng nước (m^3/s) về hồ dưới dạng sau:

$$Q = f(X) + \epsilon, \quad (1)$$

Trong đó ϵ là lỗi của mô hình. Trong biểu thức (1), ta có các biến ngẫu nhiên với M biến đầu vào $X \in \mathbb{R}^M$ và một biến đầu ra $Q \in \mathbb{R}^1$. Mục tiêu của bài toán dự báo lưu lượng nước đổ về hồ Hòa Bình trong nghiên cứu này là dùng phương pháp Deep learning tìm hàm phi tuyến f_D: $X \rightarrow Q$ để cực tiểu hóa lỗi dự báo Err (f_D) = $E_{X,Y}$ {L(Q, f_D(X))}, trong đó L(Q, f_D(X))} = (Q – f_D(X))^2 là hàm mất mát (loss function), hàm f_D(X) xây dựng được từ dữ liệu quan sát D.

Các yêu cầu phục vụ điều hành hồ chứa thường có 4 dạng dự báo: Thời gian dự báo ngắn hạn (trước 1-2 ngày), trung hạn (5-10 ngày), dài hạn (1 tháng) và siêu dài hạn (1 mùa). Dự báo ngắn hạn đòi hỏi độ chính xác cao và cấp thiết cho việc ra quyết định điều hành. Trong bài báo này, chúng tôi sử dụng phương pháp Deep learning, một trong những phương pháp phát triển rất nhanh trong thời gian gần đây và thu hút nhiều nghiên cứu trên thế giới, ứng dụng cho bài toán dự báo lưu lượng nước đổ về hồ Hòa Bình trước 10 ngày. Kết quả thực nghiệm cho thấy phương pháp Deep learning mang lại kết quả tốt hơn so với một số phương pháp học máy nổi tiếng khác như Máy véctơ hỗ trợ, rừng ngẫu nhiên, cây quyết định, k láng giềng gần nhất, LASSO khi so sánh về lỗi dự báo và độ chính xác của mô hình. Ngoài ra, việc cài đặt phân tán Deep learning giúp cải thiện tốc độ tính toán và xử lý được dữ liệu huấn luyện có dung lượng lớn. Nghiên cứu này cũng mở ra hướng ứng dụng mới cho các bài toán dự báo, giúp cải thiện độ chính xác và thời gian tính toán khi áp dụng trong lĩnh vực tài nguyên nước nói riêng và trong lĩnh vực kinh tế ở Việt Nam nói chung.

II. DEEP LEARNING

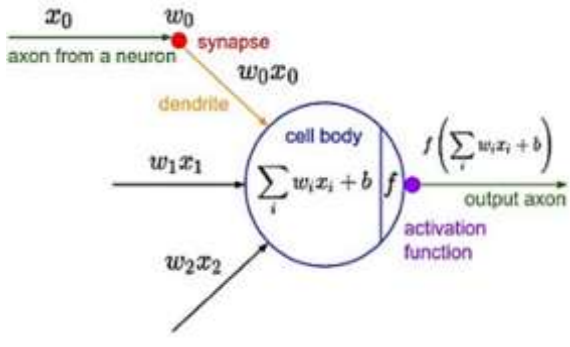
Deep learning là một tập các thuật toán học máy với ý tưởng xây dựng mô hình dữ liệu có mức độ trừu tượng cao dựa trên các dữ liệu có mức độ trừu tượng hóa thấp hơn, bằng cách phân lớp dữ liệu và các biến đổi phi tuyến [3].

Deep Neural network (DNN) là mô hình deep learning dựa trên mạng nơron nhân tạo với số lớp ẩn lớn; DNN cũng là mô hình được sử dụng với bài toán dự báo lưu lượng nước đến hồ Hòa Bình. Phần II này trình bày các khái niệm và thành phần cơ bản được sử dụng trong mô hình được đề xuất – mạng nơron sâu truyền thẳng với 3 lớp ẩn được kết nối đầy đủ.

Do đặc điểm của mô hình là mạng truyền thẳng với số nút tương đối lớn, kết nối đầy đủ nhưng tập dữ liệu huấn luyện tương đối nhỏ; một số điều chỉnh dựa trên [1] được thử nghiệm để tránh tình trạng quá khớp.

A. Nơron nhân tạo

Cấu trúc của nơron nhân tạo mô phỏng hoạt động của tế bào thần kinh trong tự nhiên được minh họa trong Hình 2; trong đó các tín hiệu kích hoạt (x_0, x_1,…) được gửi tới nơron và được điều chỉnh nhân bởi các trọng số kết nối (w_0, w_1,…). Tổng các tín hiệu này tiếp tục được điều chỉnh bởi hệ số thành kiến b (bias) – thể hiện ngưỡng lọc nội tại của tế bào. Cuối cùng, tín hiệu đầu ra của nơron được biến đổi bởi hàm kích hoạt (activation function) [5].



Hình 2. Cấu trúc một neural nhân tạo

Có nhiều lựa chọn cho các hàm kích hoạt khác nhau, tùy vào mục đích thiết kế và loại dữ liệu đầu ra mong muốn. Mô hình đề xuất trong bài báo này sử dụng hàm kích hoạt **hyperbolic tangent** *f(x) = tanh(x)* cho tất cả các nơron; hàm hyperbolic tangent và các biến thể của nó là hàm kích hoạt phù hợp với xử lý dữ liệu dạng số và là hàm kích hoạt phổ biến sử dụng cho các mô hình deep learning [4].

B. Mạng truyền thẳng kết nối đầy đủ

Các nơron kết nối với nhau tạo thành mạng nơron, có rất nhiều kiến trúc mạng khác nhau đã được thử nghiệm và sử dụng trong thực tế.

Ở kiến trúc mạng nhiều lớp truyền thẳng, các nơron đươc chia thành các nhóm hay các lớp (layer), các lớp được sắp xếp theo thứ tự tuyến tính. Các nơron trong cùng một lớp không được kết nối với nhau. Một nơron thuộc lớp trước kết nối tới các nơron thuộc lớp liền sau. Như vậy tín hiệu được truyền từ lớp đầu vào đến lớp đầu ra theo một hướng.

Việc các nơron giữa hai lớp liên tiếp được kết nối như thế nào tùy thuộc vào bài toán cụ thể và topo mạng nơron thường được lựa chọn dựa trên góc nhìn chủ quan của mô hình được đề xuất cho bài toán đó.

Trong bài toán dự báo lưu lượng nước đến hồ chứa Hòa Bình, do chưa thể biết tham số nào là quan trọng với kết quả đầu ra, mô hình đề xuất sử dụng kết nối đầy đủ: nơron thuộc lớp trước kết nối với tất cả các nơron thuộc lớp kế tiếp. Việc lựa chọn đặc trưng và tham số sẽ do thuật toán học quyết định, dựa vào việc đánh giá trọng số kết nối.

Hình 3 minh họa nơron lớp thứ i là L_{li} được kết nối đầy đủ với tất cả các nơron thuộc lớp trước và sau nó. Như vậy giữa hai lớp mạng nơron L_{l-1} và L_l chúng ta sử dụng một ma trận trọng số kí hiệu là $W^{(l-1)}$ có kích cỡ $|L_{(l-1)}| \times |L_l|$, tương tự như vậy, ma trận trọng số W^l có kích cỡ $|L_l| \times |L_{(l+1)}|$ lưu trữ các trọng số kết nối giữa các nơron thuộc lớp L_l và L_{l+1}.

C. Thuật toán lan truyền tới (feed forward)

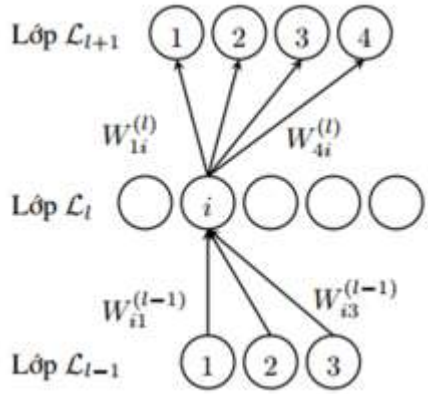
Thuật toán này sử dụng để tính toán kết quả đầu ra của mạng nơron với đầu vào là véctơ $x^{(0)}$, $W = \{W^{(0)}, W^{(1)},\ldots, W^{(L-1)}\}$ là tập hợp các ma trận trọng số, $W^{(i)}$ là ma trận trọng số của các kết nối giữa các nơron thuộc lớp i và lớp i+1.

```
function FeedForward(x^(0) ∈ R^|L0|)
       for l = 1 to L do
         z^(l) ← W^(l-1) · x^(l-1)
            x^(l) ← f(z^(l))
             end for
      return x^(L), Loss(z^(L))
        end function
```

D. Hàm mất mát (loss function)

Tổng của các độ sai lệch giữa dữ liệu ra của mạng nơron $h(x_d; W)$, và dữ liệu ra cần đạt được, $g(x_d)$, thể hiện độ tốt của tập tham số hiện tại. Nếu tập huấn luyện là cố định, tổng này về bản chất là một hàm số chỉ phụ thuộc vào tập tham số W được định nghĩa:

$$Loss(W) \equiv \sum_{d \in D} dist(h(x_d; W), g(x_d))$$



Hình 3. Kết nối đầy đủ của một neural nhân tạo

với D là tập huấn luyện, *dist* là một hàm tính độ chênh lệch giữa hai điểm dữ liệu ra [2].

Trong quá trình huấn luyện, giá trị của hàm mất mát càng nhỏ thì đầu ra của mạng nơron càng gần với đích huấn luyện. Như vậy ở góc nhìn này việc huấn luyện mạng nơron về bản chất là việc điều chỉnh tham số W để cực tiểu hóa hàm số Loss(W) [6].

E. Thuật toán huấn luyện bằng lan truyền ngược lỗi (back-propagation)

Thuật toán lan truyền ngược lỗi dựa trên ý tưởng rất đơn giản: khi đầu ra của mạng nơron không được như mong muốn, chúng ta sẽ điều chỉnh giá trị các tham số của mạng nơron. Do mạng là phân lớp và truyền thẳng, quá trình điều chỉnh sẽ đi theo chiều ngược lại, điều chỉnh từ lớp đầu ra, các lớp ẩn,… cho đến lớp đầu tiên (đây là xuất xứ của tên gọi "lan truyền ngược") [6].

```
    function BACKPROP(x^(0) ∈ R^|L0| , {W^(l)})
    Dùng thuật toán FeedForward với x^(0) để tính các giá trị z^(1),…,z^(L), x^(1),…, x^(L)
và hàm mất mát Loss(z^(L))
    δ^(L) ← ∂/∂z^(L) Loss
    for l = L − 1 to 0 do
        ∂/∂z^(l) Loss ← f′(z^(l)) ∘ (W^(l)T · δ^(l+1))
        ∂/∂W^(l) Loss ← δ^(l+1) · x^(l)T
    end for
    return  {∂/∂W^(0) Loss, · · · , ∂/∂W^(L-1) Loss }
end function
```

Các vấn đề về thuật toán được trình bày chi tiết trong [6]; đối với việc tính toán tham số điều chỉnh W, mô hình đề xuất sử dụng thuật toán SGD (Stochastic Gradient Descent), chi tiết xin xem trong [7]; thuật toán nhanh hơn so với phương pháp GD (Gradient Descent) thông thường, ngoài ra cho phép huấn luyện mạng ngay cả khi không có toàn bộ tập dữ liệu huấn luyện ngay từ đầu.

F. Khả năng xấp xỉ của mạng nơron và hiện tượng quá khớp (overfitting)

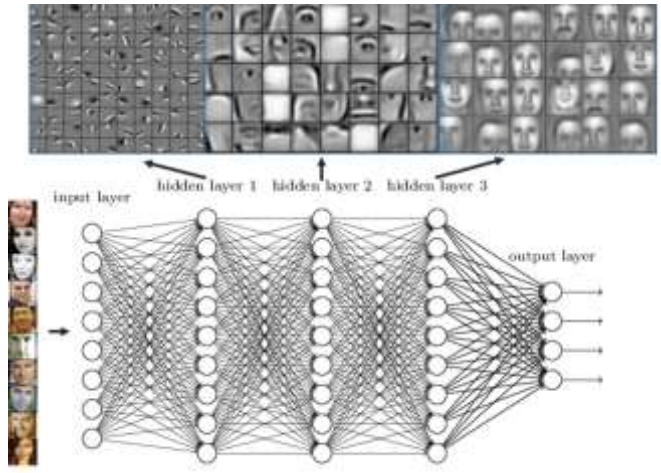
Mạng nơron đã được chứng minh có khả năng xấp xỉ vạn năng [8] với số lớp không quá lớn (4 lớp). Tuy nhiên nghiên cứu không chỉ ra được việc huấn luyện xấp xỉ như thế nào và khả năng tổng quát hóa (dự báo) của mạng. Ngoài ra, không có phương pháp nào ước lượng số nơron cần có trên mỗi lớp và topo kết nối giữa các lớp với nhau.

Khả năng xấp xỉ và mô hình hóa rất mạnh của mạng nơron không phải luôn luôn có lợi; nó dẫn đến việc mạng nơron rất dễ bị hiện tượng quá khớp (overfitting). Hiện tượng này xảy ra khi quá trình huấn luyện mạng nơron dẫn đến việc mạng đã lựa chọn xấp xỉ một hàm phức tạp quá mức cần thiết, hàm này mô phỏng hoàn hảo các tình huống huấn luyện, nhưng do cấu trúc phức tạp, hàm lại không có tính tổng quát hóa cao hoặc rất thiếu ổn định, hệ quả là mạng dự đoán không chính xác với các mẫu không có trong tập huấn luyện. Các phương pháp khắc phục hiện tượng quá khớp được trình bày trong [1], với bài toán dự báo lượng nước đổ về hồ Hòa Bình, l_1-norm [12] được lựa chọn sử dụng, lý do lựa chọn vì huấn luyện với l_1 cho tập tham số thưa giúp tăng tốc độ tính toán và giảm yêu cầu về bộ nhớ [1].

G. Deep learning

Mạng nơron được chứng minh khả năng xấp xỉ vạn năng chỉ với không quá 4 lớp, nhưng chưa có phương pháp nào cụ thể ước lượng số nơron cần thiết trên mỗi lớp. Các mô hình học sâu có kiến trúc tương tự mạng nơron nhưng dựa trên cách tiếp cận khác, với ý tưởng cơ bản là dữ liệu tại mỗi lớp sẽ có mức độ trừu tượng hóa (khái quát) cao hơn bằng cách tổ hợp các dữ liệu có mức trừu tượng hóa thấp ở lớp trước [3].

Hình 4 biểu diễn một mô hình học sâu tiêu biểu [9] sử dụng trong nhận dạng mặt người, trong đó dữ liệu đầu vào của mạng có thể là dữ liệu ở dạng thô nhất là các điểm ảnh RGB (thậm chí không cần qua tiền xử lý). Các đặc trưng được tổ hợp và tạo thành các chi tiết nhỏ ở lớp ẩn đầu tiên, sau đó tiếp tục được tái tạo và tổ hợp mức chi tiết lớn ở lớp ẩn thứ hai và cuối cùng các hình ảnh đặc trưng của toàn bộ khuôn mặt ở lớp ẩn thứ 3. Lớp đầu ra cho ra đánh giá xác suất khuôn mặt thuộc phân lớp nào (người nào).



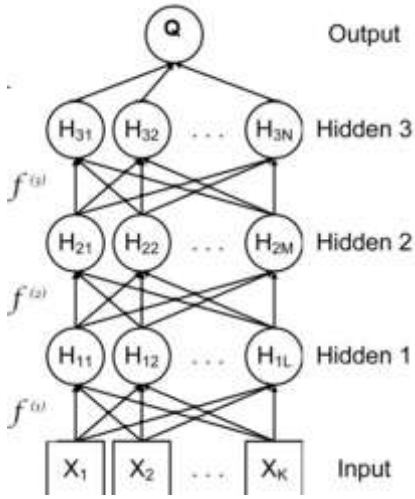
Hình 4. Mô hình học sâu trong nhận dạng mặt người

Một mô hình học sâu thường có 3 nhiệm vụ được kết hợp trong một kiến trúc mạng duy nhất:

- Các lớp đặc trưng (features): có nhiệm vụ chuyển đổi các đặc trưng thành dạng dữ liệu phù hợp để xử lý, chẳng hạn như các tầng tích chập (convolution), mẫu (subsampling), pooling,…
- Các lớp mô hình (modeling): sử dụng các thuật toán học để khái quát hóa dữ liệu, chẳng hạn nơron network, restricted BM, DBN, autoencoder,…
- Các lớp giải mã (decoding): dựa trên dữ liệu khái quát biến đổi thành đầu ra (markov random field hoặc những công cụ tương tự).

Các mạng học sâu đều có cấu trúc xác định trước, như vậy bài toán tập huấn vẫn là việc xác định giá trị các tham số trên mạng. Hiện chưa có phương pháp tập huấn nào cho phép điều chỉnh cấu trúc mạng hiệu quả [3].

H. Cấu trúc DNN thử nghiệm cho bài toán dự báo lượng nước đổ về hồ Hòa Bình



Hình 5. Cấu trúc DNN 3 lớp ẩn cho bài toán dự báo lượng nước đổ về hồ Hòa Bình

Bài toán dự báo lượng nước đổ về hồ Hòa Bình đã được thử nghiệm với nhiều mô hình dự báo như ANN (Artificial Neuron Network), SVR (Support Vector Regression), RF (random forest), LASSO [14]. Trong bài báo này, cấu trúc DNN đơn giản được đề xuất với mục đích kiểm chứng chất lượng dự báo của phương pháp deep learning so với các phương pháp đã có, kết quả thực nghiệm cho thấy chất lượng dự báo chính xác hơn so với các phương pháp trên, trong các nghiên cứu tương lai chúng tôi sẽ tiếp tục cải tiến mô hình và tiến hành thử nghiệm thực tế (đưa mô hình vào chạy thực sự với hệ thống vận hành đã có).

Mô hình thử nghiệm:

- Lớp đầu vào: dữ liệu quan sát đầu vào không qua chuẩn hóa, dạng chuỗi số liệu với cửa sổ trượt là 8 gồm cả dữ liệu về lượng mưa và lưu lượng nước đổ về hồ (xem chi tiết hơn trong phần III.A), không có thông tin về mốc thời gian (giả thiết là quãng thời gian là đều nhau). Vector dữ liệu đầu vào được lựa chọn có cấu trúc tương tự với các phương pháp đối sánh khác để đảm bảo sự khách quan và dễ dàng khi so sánh các kết quả tính toán cuối cùng.
- Các lớp đặc trưng bị loại bỏ do dữ liệu đầu vào khá đơn giản. Việc loại bỏ này cũng kiểm tra khả năng mô hình hóa dữ liệu trực tiếp từ số liệu thô của các lớp modeling. Việc DNN có kết quả tốt chưa khẳng định được việc loại bỏ các tầng features là đúng đắn, đó có thể là chủ đề của một bài báo khác.
- Các lớp ẩn: 3 lớp ẩn mỗi lớp 500 nút và kết nối đầy đủ, X là véctơ dữ liệu đầu vào và Q là biến đầu ra.
- Đầu ra: 1 đầu ra duy nhất Q là dự báo trung hạn (10 ngày), thiết kế ban đầu được dự kiến đầu ra là véctơ Q có $|Q| = 4$ gồm dự báo ngắn hạn (2 ngày), trung hạn (10 ngày), dài hạn (30 ngày) và toàn mùa, do mục tiêu chỉ có tính đánh giá mô hình và cần rút ngắn thời gian tập huấn mạng nên có điều chỉnh lại với một đầu ra trung hạn duy nhất.

- Thuật toán huấn luyện: lan truyền ngược lỗi với luật sửa lỗi hạ dốc ngẫu nhiên (stochastic gradient descent).
- Xử lý quá khớp: sử dụng phương pháp bình thường hóa tham số (regularization) norm-1 [12].

III. KẾT QUẢ THỰC NGHIỆM

A. Mô tả dữ liệu và thiết kế kịch bản dự báo

Dữ liệu dùng trong thực nghiệm được thu thập tại trạm Tạ Bú trên sông Đà từ năm 1964 đến năm 2002, đây là trạm đo lưu lượng gần hồ Hòa Bình nhất. Bộ số liệu này được đo trong mùa cạn (từ tháng 12 năm trước đến tháng 5 năm sau). Chúng tôi chia dữ liệu làm 2 tập, tập dữ liệu huấn luyện (training set) gồm 510 bản ghi đo được từ cuối năm 1964 đến đầu năm 1998, phần dữ liệu còn lại gồm 60 bản ghi dùng làm tập kiểm thử để đánh giá hiệu năng của phương pháp Deep learning.

Để mô hình hóa bài toán dự báo lưu lượng nước đổ về hồ Hòa Bình trước 10 ngày, chúng tôi đặt Q(t+10) là biến đầu ra của mô hình dự báo. Các số liệu quan trắc mưa, lưu lượng Q đo đạc được của ngày hiện tại và những ngày trước đó đều được xem xét và đưa vào mô hình. Việc lựa chọn các tiêu chí cần kinh nghiệm chuyên gia và thử nghiệm các kịch bản để lựa chọn, trong nghiên cứu này chúng tôi sử dụng kịch bản tốt nhất được trình bày trong nghiên cứu của Chen và đồng nghiệp [10]. Mối quan hệ giữa biến đầu vào và biến đầu ra của mô hình dự báo được xây dựng như sau:

$$\mathrm{Q}\,(\mathrm{t}+10) = \mathrm{f}\Big(\mathrm{Q}(\mathrm{t}), \mathrm{Q}(\mathrm{t}-10),\ \mathrm{Q}(\mathrm{t}-20),\ \mathrm{X}(\mathrm{t}), \mathrm{X}(\mathrm{t}-10), \mathrm{X}(\mathrm{t}-20), \mathrm{X}_{day}(\mathrm{t}), \mathrm{Q}_{day}(\mathrm{t})\Big), \quad (2)$$

trong đó $\mathrm{Q}(\mathrm{t}), \mathrm{Q}(\mathrm{t}-10),\ \mathrm{Q}(\mathrm{t}-20)$ là lưu lượng nước về hồ ở thời điểm hiện tại, thời điểm trước đó 10 và 20 ngày tương ứng. Lượng mưa quan trắc ở thời điểm hiện tại, thời điểm trước đó 10 và 20 ngày được lưu trữ tương ứng trong các biến $\mathrm{X}(\mathrm{t}), \mathrm{X}(\mathrm{t}-10)$ và $\mathrm{X}(\mathrm{t}-20)$. Hai biến $\mathrm{X}_{day}(\mathrm{t})$ và $\mathrm{Q}_{day}(\mathrm{t})$ chỉ số liệu mưa quan trắc và lưu lượng nước về hồ của ngày hiện tại. Lý do chọn các yếu tố này để xây dựng mô hình dự báo vì trong nghiên cứu thủy văn, lượng mưa tại lưu vực luôn ảnh hưởng nhiều đến lượng nước chảy về hồ chứa. Hàm quan hệ $\mathrm{f}(\cdot)$ được xây dựng từ dữ liệu sử dụng phương pháp Deep learning. Lưu lượng Q được đo theo m^3/s và số liệu mưa quan trắc được đo theo đơn vị *mm*.

B. Phương pháp đánh giá

Chúng tôi dùng căn bình phương sai số (Root mean squared error-RMSE) và hệ số xác định bội (coefficient of determination) R^2 để đánh giá tính hiệu quả của các mô hình dự báo dựa trên phương pháp học máy:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Q_i - \hat{Q}_i)^2} \text{ và} \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(Q_i - \hat{Q}_i)}{\sum_{i=1}^{N}(Q_i - \bar{Q}_i)} \quad (4)$$

Trong đó: Q_i, $\hat{Q}_i$ và $\bar{Q}$ chỉ giá trị lưu lượng đo đạc được, giá trị dự đoán và giá trị trung bình của mẫu thứ i tương ứng. Mô hình dự báo cho kết quả tốt là mô hình đạt được sai số RMSE nhỏ và R^2 lớn. Hệ số xác định bội R^2 cao là một dấu hiệu cho thấy mối liên hệ giữa các yếu tố đầu vào và biến lưu lượng Q chặt chẽ. Giá trị R^2 càng cao cho thấy mô hình sử dụng để dự báo có khả năng giải thích càng tốt các thay đổi của lưu lượng nước đổ về hồ giữa các yếu tố mưa quan trắc và lưu lượng đo được trong quá khứ.

C. Kết quả dự báo lưu lượng nước đổ về hồ Hòa Bình trước 10 ngày

Thực nghiệm được tiến hành trên môi trường R (https://cran.r-project.org/), chúng tôi sử dụng gói h2o [11] để xây dựng mô hình dự báo dùng Deep learning. Phần cấu trúc của mạng chúng tôi đã nêu ở mục II và Hình 5, tham số phạt $\ell_1=10^{-6}$ và tỷ lệ học rho=0.999, số lần lặp dữ liệu huấn luyện epochs = 10. Để kiểm nghiệm tính hiệu quả của phương pháp nghiên cứu, chúng tôi thử nghiệm các thuật toán học máy nổi tiếng khác như Rừng ngẫu nhiên (RF), Máy véctơ hỗ trợ (SVM), k láng giềng (KNN), cây hồi quy, LASSO; các mô hình này được thực nghiệm dùng qua giao diện gói phần mềm caret [13]. Khi xây dựng mô hình dự báo, chúng tôi sử dụng kỹ thuật kiểm tra chéo 10-fold để lựa chọn mô hình và tìm tham số tối ưu của từng mô hình. Các thực nghiệm được tiến hành trên máy tính sử dụng hệ điều hành Windows Server 2012 64-bit, có cấu hình IntelR XeonR CPU E5-2640 2.5 GHz, 24 cores, 8 MB cache và 128 GB RAM. Phương pháp Deep learning được cài đặt phân tán và các mô hình học máy khác đều được cài đặt song song sử dụng hết 24 cores trên để huấn luyện, tìm tham số tối ưu và các thực nghiệm khác.

Bảng 1 trình bày kết quả dự báo của phương pháp Deep Learning so sánh với các giải thuật học máy nổi tiếng khác. Ở 2 cột R^2 và RMSE kết quả dự đoán với R^2 cao nhất và lỗi dự đoán thấp nhất được in đậm và gạch dưới.

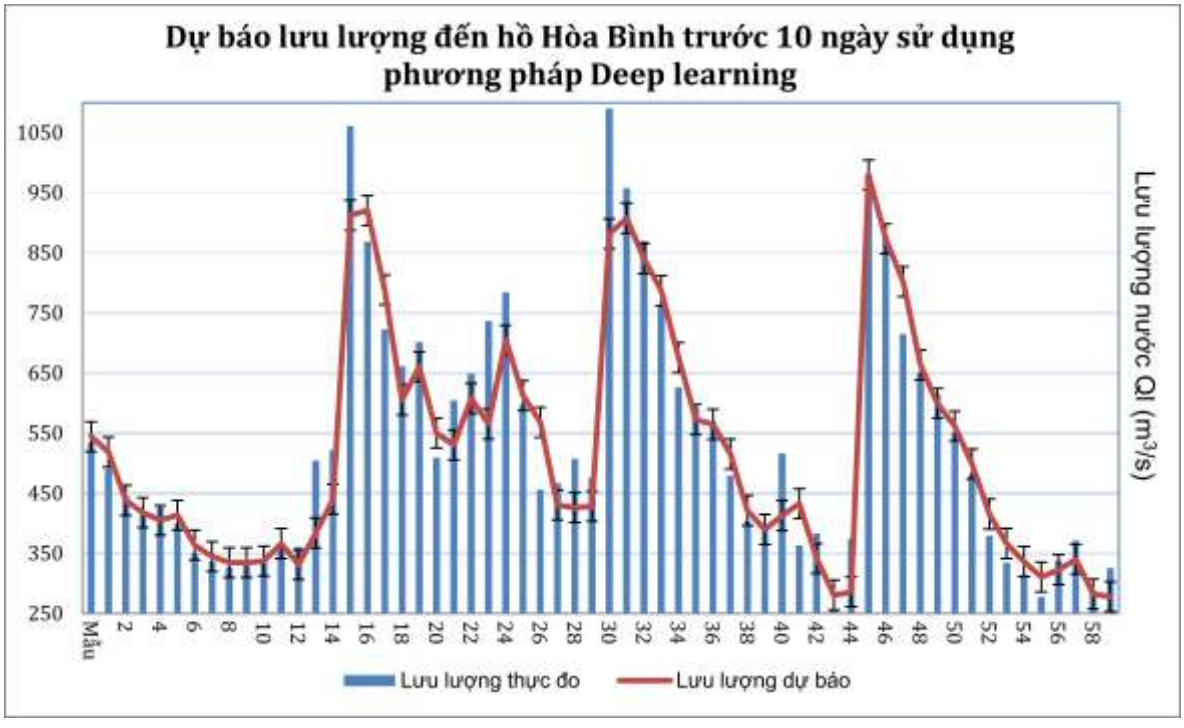
Ta có thể thấy nhận thấy kết quả dự báo do phương pháp Deep learning mang lại có kết quả tốt nhất, hệ số xác định bội R^2 đạt 92%, điều này phản ảnh mô hình giải thích tốt sự khác biệt các số liệu mưa quan trắc, lưu lượng đo đạc (biến đầu vào) với sự thay đổi về lưu lượng nước đổ về hồ. Cấu trúc mạng cài đặt trong thực nghiệm này có 3 lớp ẩn, mỗi lớp có 500 nơron kết nối đầy đủ và dùng hàm kích hoạt tanh(x) cho tất cả các nơron trong mạng. Các mô hình LASSO và Rừng ngẫu nhiên cũng cho kết quả ấn tượng với R^2 đạt gần 90% và sai số dự báo tương ứng là 66.3 m^3/s và 68.5 m^3/s.

Phương pháp LASSO đạt lỗi dự báo chỉ sau Deep learning, đây là phương pháp cải tiến từ mô hình hồi quy tuyến tính có đưa thêm hàm phạt để giảm lỗi của mô hình [12], LASSO là phương pháp hồi quy tuyến tính khá hiệu quả và được ứng dụng nhiều trong thực tế bởi tính dễ diễn giải và chất lượng dự báo cao.

Bảng 1. Kết quả dự báo của phương pháp Deep learning trên dữ liệu kiểm thử

TT	Mô hình dự báo	Tham số tối ưu	R^2	RMSE (m^3/s)
1	LASSO	λ= 0.9	0.89	66.3
2	K láng giềng (KNN)	k = 16	0.87	87.9
3	Cây hồi quy (CART)	Complexity parameter (cp)=0.0022	0.82	89.2
4	Máy véctơ hỗ trợ (SVR)	Nhân Radial: σ = 0.545 và C = 2	0.73	119.9
5	Rừng ngẫu nhiên (RF)	mtry = 3 và K=1000	0.89	68.5
6	Deep Learning	Ba lớp ẩn kết nối đầy đủ với 500 nơron mỗi lớp; Rho=0.99, ℓ_1=10^{-6}, epochs = 10	**0.92**	**60.37**

Hình 6 hiển thị đường dự báo lưu lượng nước đổ về hồ Hòa Bình trước 10 ngày, các cột lưu lượng thực đo hiển thị màu xanh trên biểu đồ, đường dự báo hiển thị với màu nâu đỏ. Trên mỗi điểm dự báo có các đồ thị nhỏ dạng thanh ngang biểu thị độ lệch chuẩn về lỗi dự báo. Ta có thể thấy rằng các giá trị lưu lượng dự báo bám khá sát với giá trị thực đo, đặc biệt là các mẫu thử nghiệm từ 1 đến 12, từ 44 đến 54. Tại một số điểm có lưu lượng Q khá lớn (trên 1,050 m^3/s) như mẫu thứ 15 và mẫu thứ 30, phương pháp Deep learning cho kết chưa được khả quan. Tuy nhiên đây là những mẫu bất thường và tồn tại không nhiều trong tập dữ liệu thử nghiệm. Nhìn chung, chất lượng dự báo của phương pháp nghiên cứu là khả quan dựa trên giá trị R^2 và đường dự báo trực quan trong Hình 6 bám khá sát với giá trị lưu lượng thực đo.



Hình 6. Kết quả dự báo lưu lượng nước về hồ Hòa Bình khi so sánh với giá trị lưu lượng thực đo

IV. KẾT LUẬN

Chúng tôi đã trình bày phương pháp Deep learning ứng dụng cho bài toán dự báo trung hạn lưu lượng nước về hồ Hòa Bình trước 10 ngày. Đóng góp chính của bài báo là nghiên cứu phương pháp Deep learning ứng dụng vào một bài toán thực tế, mô hình nghiên cứu Deep learning trong bài báo này là một phương pháp thuộc lĩnh vực đang phát triển rất nhanh và nhận được nhiều sự quan tâm nghiên cứu trên thế giới trong thời gian gần đây. Các phương pháp học máy nổi tiếng khác như LASSO, cây hồi quy, k láng giềng, véctơ hỗ trợ hồi quy, rừng ngẫu nhiên đã được nghiên cứu, cài đặt thử nghiệm và so sánh với Deep learning để đánh giá hiệu năng của các mô hình dự báo. Kết quả thực nghiệm cho thấy phương pháp Deep learning cho kết quả dự báo tốt nhất với R^2 đạt 92% và vượt trội so với các phương pháp khác khi so sánh dựa trên R^2 và RMSE. Trong tương lai, chúng tôi sẽ áp dụng kết quả nghiên cứu mở rộng cho các bài toán kinh tế và những bài toán liên quan đến dự đoán với số chiều cao ở Việt Nam.

V. LỜI CẢM ƠN

Bài báo được hỗ trợ bởi đề tài "Ứng dụng Công nghệ thông tin xây dựng mô hình hỗ trợ công tác dự báo lũ trên sông và cảnh báo ngập lụt lưu vực sông Tích - sông Bùi" thuộc Sở KHCN Thành phố Hà Nội. Nhóm tác giả xin cảm ơn GS. Hà Văn Khối, Khoa Thủy văn và Tài nguyên nước, Trường Đại học Thuỷ lợi đã hỗ trợ cung cấp số liệu thử nghiệm.

TÀI LIỆU THAM KHẢO

[1]. G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. "Improving Neural networks by preventing co-adaptation of feature detectors." Technical Report, arXiv:1207.0580, pp. 1–18, 2012.
[2]. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio. "Maxout networks". arXiv preprint arXiv:1302.4389, 2013
[3]. Bengio, Yoshua. "Learning Deep Architectures for AI" . Foundations and Trends in Machine Learning: Vol. 2: No. 1, pp 1–127, (2009)
[4]. LeCun, Yann; Bengio, Yoshua; Hinton, Geoffrey (2015). "Deep learning". Nature 521: pp 436–444
[5]. F. Rosenblatt. "The perceptron, a perceiving and recognizing automaton". Project Para. Cornell Aeronautical Laboratory, 1957
[6]. Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986), "Learning representations by back-propagating errors". Nature 323, pp 533-536
[7]. Léon Bottou. "Large-Scale Machine Learning with Stochastic Gradient Descent", Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010), 177–187, Edited by Yves Lechevallier and Gilbert Saporta, Paris, France, August 2010, Springer.
[8]. Haykin, Simon (1998). "Neural Networks: A Comprehensive Foundation", Volume 2, Prentice Hall. ISBN 0-13-273350-1.
[9]. Honglak Lee, Roger Grosse, Rajesh Ranganath and Andrew Y. Ng, "Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations", ICML 2009.
[10]. Chen, Jeng-Fung, Ho-Nien Hsieh, and Quang Hung Do. "Forecasting Hoabinh Reservoir's Incoming Flow: An Application of Neural Networks with the Cuckoo Search Algorithm." Information 5.4 (2014): 570-586.
[11]. Candel, Arno, Viraj Parmar, Erin LeDell, and Anisha Arora. "Deep Learning with H2O." (2015).
[12]. Tibshirani, R., Bickel, P., Ritov, Y. and Tsybakov, A., 1996. Least absolute shrinkage and selection operator. Software: http://www-stat.stanford.edu/~tibs/lasso.html.
[13]. Max Kuhn. Building predictive models in r using the caret package. *Journal of Statistical Software*, 28(5):1–26, 2008.
[14]. Nguyen, Thanh-Tung. "An ℓ_1-Regression Random Forests Method for Forecasting of Hoa Binh Reservoir's Incoming Flow." In Knowledge and Systems Engineering (KSE), 2015 Seventh International Conference on, pp. 360-364. IEEE, 2015.
[15]. Kumar, K.V. Neural Network Prediction of Interfacial Tension at Crystal/Solution Interface. Ind. Eng. Chem. Res. 2009, 48, 4160–4164.
[16]. Coulibaly, P.; Anctil, F.; Bobée, B. Multivariate Reservoir Inflow Forecasting Using Temporal Neural Networks. J. Hydrol. Eng. 2001, 6, 367–376.
[17]. Roosta, A.; Setoodeh, P.; Jahanmiri, A. Artificial Neural Network Modeling of Surface Tension for Pure Organic Compounds. Ind. Eng. Chem. Res. 2011, 51, 561–566.
[18]. Zendehboudi, S.; Ahmadi, M.A.; Mohammadzadeh, O.; Bahadori, A.; Chatzis, I. Thermodynamic Investigation of Asphaltene Precipitation during Primary Oil Production: Laboratory and Smart Technique. Ind. Eng. Chem. Res. 2013, 52, 6009–6031.
[19]. Coulibaly, P.; Anctil, F.; Bobée, B. Multivariate Reservoir Inflow Forecasting Using Temporal Neural Networks. J. Hydrol. Eng. 2001, 6, 367–376.
[20]. Nayak, P., et al. (2005) Short-term flood forecasting with a neurofuzzy model, Water Resources Research, 41.
[21]. Shamseldin, A. (2010) Artificial neural network model for river flow forecasting in a developing country, Journal of Hydroinformatics, 12, 22-35.
[22]. Shamseldin, A.Y. (1997) Application of a neural network technique to rainfall-runoff modelling, Journal of Hydrology, 199, 272-294.
[23]. Shamseldin, A.Y. and O'CONNOR, K.M. (1999) A real-time combination method for the outputs of different rainfall-runoff models, Hydrological Sciences Journal, 44, 895-912.
[24]. Shamseldin, A.Y. and O'Connor, K.M. (2001) A non-linear neural network technique for updating of river flow forecasts, Hydrology and Earth System Sciences Discussions, 5, 577-598.

DEEP LEARNING: AN APPLICATION TO HOABINH RESERVOIR'S INCOMING FLOW FORECAST

Truong Xuan Nam, Tung Nguyen

ABSTRACT*—The prediction of the incoming flow plays an important role for the operation of the Hoa Binh reservoir as well as prevention of natural hazards in Vietnam. In this paper, we apply a Deep learning method to forecast the incoming flows on HoaBinh Reservoir in the period of 10 days. Experimental results showed that our predicted model provided positive results and outperformed some well-known machine learning methods such as LASSO, KNN, CART, Support Vector Machine, Random Forests; the coefficient of determination R^2 of Deep learning achieves 92%. We finally concluded that the Deep learning method is useful and feasible and opens a new direction for solving time series problems.*

Keywords*—Deep learning, LASSO, k nearest neighbors, SVR, random forests, data mining, machine learning.*