

XÂY DỰNG VÀ ĐÁNH GIÁ CƠ SỞ DỮ LIỆU VỀ DỊCH BỆNH COVID-19 TẠI VIỆT NAM BẰNG MẠNG NEURON NHÂN TẠO

FORMATION AND EVALUATION OF A DATABASE OF COVID-19 DISEASES IN VIETNAM USING AN ARTIFICIAL NEURAL NETWORK

Huỳnh Lương Nghĩa, Đinh Văn Quang, Bùi Văn Tấn, Nguyễn Thị Thủy

Trường Đại học Điện lực

Ngày nhận bài: 28/2/2023 Ngày chấp nhận đăng: 30/5/2023, Phản biện: TS Đặng Thúy Hằng

Tóm tắt:

Bài báo trình bày cách tạo cơ sở dữ liệu về bệnh Covid-19 tại Bệnh viện Bệnh nhiệt đới Trung ương (gọi là cơ sở dữ liệu CovidVN) và sau đó xây dựng mạng neuron học tập dựa trên cơ sở dữ liệu này để hỗ trợ chẩn đoán bệnh.

Cơ sở dữ liệu CovidVN được xây dựng trên cơ sở xử lý và tổng hợp kết quả xét nghiệm chẩn đoán thực tế của bệnh nhân mắc Covid-19 với số lượng mẫu lớn và phù hợp với cấu trúc cơ sở dữ liệu bệnh Covid-19 của Hệ thống Y tế Israel (gọi tắt là Cơ sở dữ liệu COVIDIsr).

Sau đó mạng neuron nhân tạo MLP tương ứng với hai cơ sở dữ liệu này sẽ được phát triển bằng hộp công cụ học sâu của phần mềm MATLAB; kết quả luyện các mạng này và độ chính xác của chúng được so sánh với nhau để đánh giá tương đối chất lượng của CSDL CovidVN.

Từ khóa:

Mạng neuron nhân tạo, cơ sở dữ liệu virus corona (COVID-19), perceptron đa lớp (MLP).

Abstract:

In this article, we present how to create a database of Covid-19 diseases at National Hospital of Tropical Diseases (called the CovidVN database) and then develop a learning neural network based on this database to diagnose this disease.

The CovidVN Database is built based on the processing of real diagnostic test results of Covid-19 patients with a large number of samples and in accordance with the structure of the Israeli Health System Covid-19 disease database (called COVIDIsr Database).

Then two MLP Artificial Neural Networks corresponding to these two databases will be developed using the Deep Learning Toolbox of MATLAB software; the results of training these networks and their accuracy are compared with each other to assess the relative quality of the CovidVN database.

Keywords:

Artificial neuron networks (ANN), Covid-19 flu database (COVID-19); MultiLayer Perceptron (MLP).

1. GIỚI THIỆU CHUNG

Việc ứng dụng công nghệ trí tuệ nhân tạo

vào chẩn đoán và điều trị y khoa đang bùng nổ trên thế giới [1] và bắt đầu ở Việt

Nam [2], [3], [4]..., trong đó có ứng dụng [5] sử dụng mạng nơron nhận dạng hình ảnh để chẩn đoán Covid. Tuy nhiên, điều kiện tiên quyết để áp dụng phương pháp học máy trong chẩn đoán y khoa là phải có cơ sở dữ liệu về chẩn đoán bệnh lý..., điều mà trên thế giới chỉ một số nước xây dựng và công bố như Mỹ, Israel..., nhưng đối với Việt Nam thì hoàn toàn chưa có.

Vì vậy, để ứng dụng các phương pháp học máy (công nghệ trí tuệ nhân tạo) nhằm xây dựng hệ thống hỗ trợ chẩn đoán bệnh Covid-19, trong công trình này, trước hết chúng tôi sẽ thiết kế cơ sở dữ liệu dùng để chẩn đoán bệnh Covid-19 (gọi tắt là cơ sở dữ liệu COVID19VN); sau đó dựa trên cơ sở dữ liệu này sẽ xây dựng một mạng nơron nhân tạo kiểu MLP để hỗ trợ chẩn đoán bệnh Covid. Hiệu suất chất lượng của mạng nơron này sẽ được so sánh với mạng nơron có cùng thiết kế nhưng dựa trên Bộ dữ liệu nghiên cứu mở về Covid-19 (CORD-19) của Bộ Y tế Israel [6].

2. DỮ LIỆU VÀ PHƯƠNG PHÁP

2.1. Dữ liệu

2.1.1. Cơ sở dữ liệu CORD-19

Cơ sở dữ liệu CORD-19 được lấy từ Bộ dữ liệu nghiên cứu mở cho AI của Đại học Allen (CORD-19) hợp tác với Sáng kiến Chan Zuckerberg, Trung tâm Công nghệ An ninh và Nội trợ của Đại học Georgetown, Nghiên cứu của Microsoft, IBM và Thư viện Y khoa Quốc gia - Viện Y học Quốc gia Health, phối hợp với Văn phòng Chính sách Khoa học và Công

nghệ của Nhà Trắng [5]. Cơ sở dữ liệu này chứa các thông số đặc trưng thu được từ 578 bệnh nhân và 6 thuộc tính đầu vào như: giới tính, ho, sốt, viêm họng, khó thở, đau đầu liên quan đến 3 đầu ra tương ứng với 3 kết quả chẩn đoán bằng xét nghiệm test nhanh kháng nguyên SARS-CoV: âm tính, dương tính với cúm Covid-19 và khác.

2.1.2. Cơ sở dữ liệu COVID19 VN

Dựa trên cấu trúc thuộc tính của cơ sở dữ liệu Open Research Dataset CORD-19, có thể thiết kế cơ sở dữ liệu COVID19VN với 6 thuộc tính (giới tính, ho, sốt, đau họng, khó thở, đau đầu), cỡ mẫu lớn hơn 500, theo biểu mẫu thu thập dữ liệu tương đương... Việc triển khai thiết kế này được thực hiện tại Bệnh viện Bệnh Nhiệt đới Trung ương Hà Nội - bệnh viện tuyến đầu trong công tác phòng chống các bệnh truyền nhiễm và nhiệt đới - trong đợt cao điểm của dịch Covid-19 đầu năm 2022 với sự tham gia của các bác sĩ, điều dưỡng... về nhân viên và bệnh nhân đã từng đến khám tại các phòng khám chuyên khoa của bệnh viện... Kết quả là đã xây dựng được cơ sở dữ liệu đầu tiên về bệnh cúm Covid-19 cho bệnh nhân Việt Nam (COVID19VN) (Bảng 1). Nó bao gồm các thông số đặc trưng thu được từ 504 bệnh nhân và 6 thuộc tính đầu vào được đề cập ở trên và được liên kết với 3 đầu ra: âm tính, dương tính với cúm Covid-19 và khác (được xác định bằng xét nghiệm test nhanh kháng nguyên SARS-CoV).

2.1.3. Tiền xử lý cơ sở dữ liệu; lọc các thuộc tính tổn thất lớn, chuẩn hóa

Trước khi được sử dụng để huấn luyện mạng nơron, cơ sở dữ liệu cần được xử lý trước để lọc ra các thuộc tính có tổn thất lớn và sau đó giá trị của các thuộc tính còn lại được chuẩn hóa sao cho chúng nằm trong khoảng (0,1).

Như vậy, sau khi xử lý sơ bộ, cơ sở dữ liệu CORD-19 dùng để huấn luyện mạng nơron học máy bao gồm 2 ma trận: ma trận CORD-19Inputs (IP1) có kích thước

6×578 tương ứng với 6 thuộc tính và 578 mẫu và ma trận CORD-19Targets (OP1) có kích thước 3×578 tương ứng với 3 kết quả chẩn đoán bệnh và 578 mẫu.

Tương tự, cơ sở dữ liệu COVID19VN sau khi xử lý sơ bộ gồm 2 ma trận: ma trận COVID19VNInputs (IP2) có kích thước 6×504 tương ứng với 6 thuộc tính và 504 mẫu và ma trận COVID19VNTargets (OP2) có kích thước 3×504 tương ứng với 3 kết quả chẩn đoán bệnh và 504 mẫu bệnh phẩm.

Bảng 1. Cấp độ cấu trúc của cơ thể và phương pháp chẩn đoán

	Attributes - symptoms (Network input)						Diagnostic results (Network output)		
Sample number	Gender	cough	fever	sore throat	dyspnea	headache	negative	positive	other
1	0	0	0	0	0	0	1	0	0
2	0	0	0	0	0	0	0	1	0
3	0	0	0	0	0	0	1	0	0
4	1	0	0	0	0	0	1	0	0
5	0	0	0	0	0	0	1	0	0
6	0	0	0	0	0	0	1	0	0
7	1	0	0	0	0	0	1	0	0
8	0	0	0	0	0	0	1	0	0
9	0	0	0	0	0	0	1	0	0
10	1	0	0	0	0	0	1	0	0
11	1	0	0	0	0	0	1	0	0
12	1	0	0	0	0	0	1	0	0
13	1	0	0	0	0	0	1	0	0
14	0	0	0	0	0	0	1	0	0
15	1	0	0	0	0	0	1	0	0
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
500	0	0	0	0	0	1	0	1	0

	Attributes - symptoms (Network input)						Diagnostic results (Network output)		
Sample number	Gender	cough	fever	sore throat	dyspnea	headache	negative	positive	other
501	0	0	0	1	0	0	0	1	0
502	0	0	0	0	0	0	0	1	0
503	1	0	0	0	0	0	0	0	1
504	1	0	0	0	0	0	0	0	1

Note:

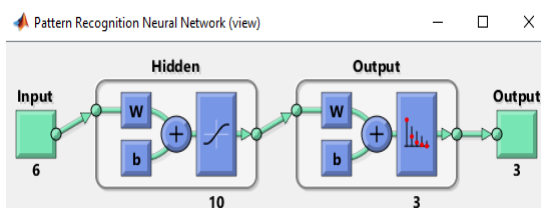
- Yes symptoms – 1 ;
- No symptoms - 0.

2.2. Phương pháp

Từ quan điểm của lý thuyết nhận dạng, chẩn đoán y tế thuộc về bài toán phân lớp thường được giải quyết bằng cách áp dụng mạng nơron nhân tạo Perceptron đa lớp (MLP) [2], [3], [4]. Như vậy, đặt ra yêu cầu thiết kế 2 mạng đa lớp trực tiếp có cùng cấu trúc dựa trên 2 cơ sở dữ liệu CORD-19 và COVID19VN.

2.2.1. Thiết kế mạng đa lớp

Để đảm bảo tính khoa học và thuận tiện trong việc xây dựng và khảo sát mạng, phần mềm Neuron Network Toolbox trong phần mềm Matlab [7] sẽ được sử dụng để thiết kế mạng. Như vậy, dựa trên cấu trúc đầu ra/đầu vào của các cơ sở dữ liệu nói trên, có thể đề xuất thiết kế mạng với số lớp và số nơron cụ thể như sau:



Hình 1. Mạng Neuron MLP

Cụ thể, mạng này có đầu vào là 6 nơron tương ứng với 6 thuộc tính, lớp ẩn sigmoid gồm 10 nơron và lớp đầu ra có 3 nơron là 3 đầu ra tương ứng với 3 kết quả của bệnh được chẩn đoán.

2.2.2. Thuật toán huấn luyện mạng

Để tối ưu hóa/huấn luyện mạng MLP, phương pháp phổ biến nhất hiện nay là sử dụng thuật toán lan truyền ngược. Thuật toán này hội tụ đến một giải pháp giảm thiểu lỗi bình phương trung bình vì việc hiệu chỉnh trọng số và sai lệch của thuật toán được thực hiện theo hướng ngược lại với vector gradient của hàm lỗi bình phương trung bình có trọng số.

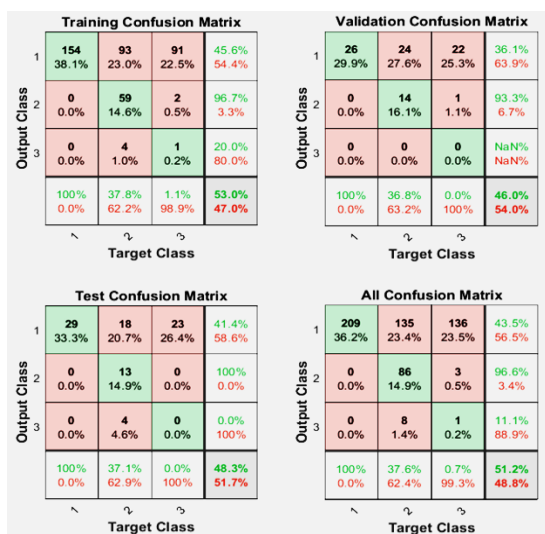
2.2.3. Phân bố tỷ lệ mẫu

Việc đào tạo, kiểm tra và xác nhận mạng được thực hiện với tỷ lệ mẫu lần lượt là 70%, 15% và 15%.

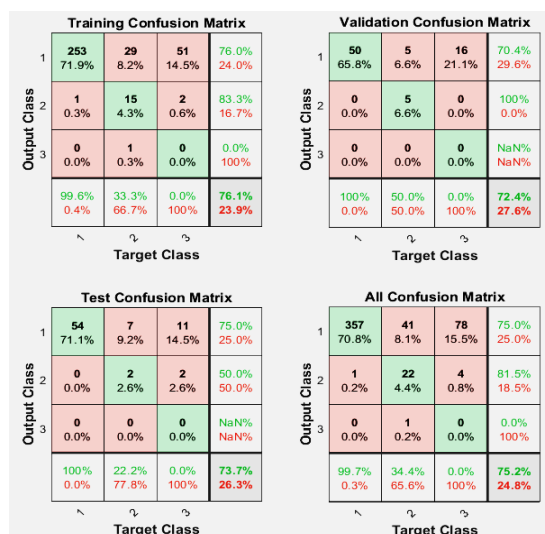
3. KẾT QUẢ, ĐÁNH GIÁ

3.1. Kết quả

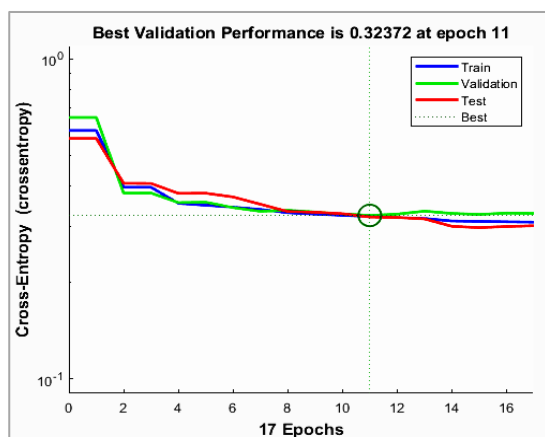
Kết quả đào tạo mạng với cơ sở dữ liệu CORD-19 được hiển thị trong Hình 2 và Hình 3.



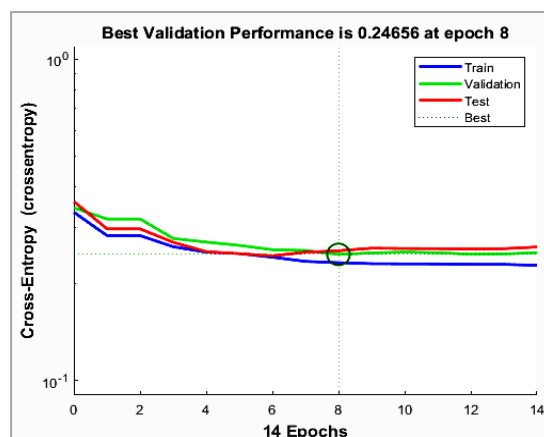
Hình 2. Ma trận kết quả đào tạo, xác nhận, kiểm tra và lỗi tổng thể của mạng dựa trên cơ sở dữ liệu CORD-19



Hình 4. Ma trận kết quả đào tạo, xác nhận, kiểm tra và lỗi tổng thể của mạng dựa trên cơ sở dữ liệu COVID19VN



Hình 3. Hiệu suất xác thực tốt nhất của mạng dựa trên cơ sở dữ liệu CORD-19



Hình 5. Hiệu suất xác thực tốt nhất của mạng dựa trên cơ sở dữ liệu COVID19VN

Kết quả đào tạo mạng với cơ sở dữ liệu CORD-19 được hiển thị trong Hình 4 và Hình 5.

Bảng 2. Kết quả chính của bài kiểm tra đánh giá mạng nơron đối với các tùy chọn chỉ định trọng lượng- Cơ sở dữ liệu COVIDIsr (CORD-19)

Trường hợp	Cơ sở dữ liệu	Tỷ lệ thành công [%]	Tỷ lệ thất bại [%]	Hiệu suất xác thực tốt nhất	Chu kỳ
1	CORD-19	51.2	48.8	0.32372	11
2	COVID19VN	75.2	24.8	0.24656	8

3.2. Đánh giá

▪ Do dữ liệu đầu vào giữa cơ sở dữ liệu CORD-19 và cơ sở dữ liệu COVID19VN có sự khác nhau nên kết quả thu được như sau:

- Tỷ lệ thành công đối với cơ sở dữ liệu COVID19VN lên tới 72,4% và đối với cơ sở dữ liệu COVIDIsr (CORD-19) chỉ là 51,2%;

- Khoảng thời gian hội tụ đối với cơ sở dữ liệu COVID19VN là 8 chu kỳ trong khi đối với cơ sở dữ liệu CORD-19 là 11 chu kỳ;

- Kết quả xác thực tốt nhất cho cơ sở dữ liệu COVID19VN là 0,24656 trong khi đối với cơ sở dữ liệu CORD-19 là 0,32372.

▪ Kết quả này cho thấy hiệu suất của mạng dựa trên cơ sở dữ liệu COVID19VN cao hơn so với mạng dựa trên cơ sở dữ liệu CORD-19. Sự khác biệt về hiệu quả chẩn đoán của hai mạng này có thể là do sự khác biệt về độ chính xác trong cách thức/phương pháp kiểm tra kết quả chẩn đoán.

▪ Tỷ lệ chẩn đoán thành công thấp của cả hai mạng (< 75%) có thể là do việc sử

dụng không phù hợp các thuộc tính có trọng số không cao tương ứng với các mức cấu trúc cơ thể vĩ mô...

4. KẾT LUẬN

Nghiên cứu này giới thiệu cơ sở dữ liệu mới về bệnh Coronavirus COVID-19 (COVID19VN) được xây dựng tại Bệnh viện Bệnh Nhiệt đới Trung ương Hà Nội. Ngoài ra đã xây dựng hai mạng nơron nhân tạo MLP hỗ trợ chẩn đoán bệnh COVID-19 dựa trên cơ sở dữ liệu này và cơ sở dữ liệu CORD-19 của Israel.

Hiệu quả cao hơn của mạng xây dựng cơ sở dữ liệu COVID19VN so với mạng dựa trên cơ sở dữ liệu CORD-19 cho phép khẳng định chất lượng của cơ sở dữ liệu COVID19VN cũng như mạng nơron học máy dựa trên nó. Do đó, chắc chắn có thể sử dụng mạng nơron được thiết kế trên trong việc hỗ trợ chẩn đoán bệnh Covid-19 trong điều kiện Việt Nam.

Nghiên cứu này cũng chỉ ra hướng phát triển tiếp theo trong việc ứng dụng trí tuệ nhân tạo trong y tế tại Việt Nam, đó là tạo ra cơ sở dữ liệu chẩn đoán bệnh nói chung và bệnh cúm COVID-19 nói riêng với cấu trúc thuộc tính hợp lý hơn tương ứng với các cấp độ cấu trúc vi mô của cơ thể.

TÀI LIỆU THAM KHẢO

- [1] S. Kaur et al., "Medical Diagnostic Systems Using Artificial Intelligence (AI) Algorithms: Principles and Perspectives," in IEEE Access, vol. 8, pp. 228049-228069, 2020, doi: 10.1109/ACCESS.2020.3042273.
- [2] Huynh Luong Nghia - Cong Doan Truong - Dinh Van Quang - Xuan Thu Do - Anh Ngoc Le. An Application of a Method of Weighting Assigning for Attributes Based on the Level of Different Body Structures to Improve the Artificial Neural Network for Diagnosing Hepatitis. Lecture Notes in Networks and Systems - ICISN 2021, (ISSN: 2367-3370).

- [3] H.L. Nghia - N.T. Thuy and D.V. Quang. Pathological diagnosis Neuron Network with inputs corresponding with structure levels of the body. Tạp chí nguyên cứu khoa học và công nghệ quân sự, số 57A, 11-2018.
- [4] H.L. Nghia - D.V. Quang - D.T.B Ngọc and N.T. Thuy. Performance improvement of diag-nostic neuron network for arrhythmia by method of assigning weights to body structure levels. Tạp chí khoa học và công nghệ năng lượng - trường đại học điện lực (ISSN: 1859 - 4557).
- [5] G. Jain, D. Mittal, D. Thakur, and M.K. Mittal, "A deep learning approach to detect Covid-19 coronavirus with x-ray images" Biocybernetics and Biomedical Engineering, vol. 40, no. 4, pp. 1391–1405, 2020.
- [6] <https://cset.georgetown.edu/publication/Covid-19-open-research-dataset-cord-19/>
- [7] <https://www.mathworks.com/products/matlab-online.html>, last visit: 5/14/18

Giới thiệu tác giả:



Tác giả Huỳnh Lương Nghĩa tốt nghiệp đại học tại Trường Kỹ sư Tên lửa phòng không Odetxa (Liên Xô) năm 1973; nhận bằng Tiến sĩ ngành điều khiển học tại Minsk, Belarus năm 1999; nguyên Chủ nhiệm Bộ môn Điện tử Y Sinh, Học viện Kỹ thuật quân sự (2001-2008); được phong hàm Phó Giáo sư ngành tự động hóa năm 2006. Hiện nay tác giả là giảng viên Khoa Điện tử Viễn thông - Trường Đại học Điện lực.

Lĩnh vực nghiên cứu: ứng dụng trí tuệ nhân tạo trong chẩn đoán và điều trị y tế.



Tác giả Nguyễn Thị Thủy tốt nghiệp đại học tại Trường Đại học Sư phạm I Hà Nội năm 1998; nhận bằng Thạc sĩ ngành công nghệ vật liệu điện tử năm 2003, bằng Tiến sĩ ngành công nghệ vật liệu điện tử năm 2014 tại Đại học Bách khoa Hà Nội. Hiện nay tác giả là giảng viên chính Khoa Điện tử Viễn thông, Trường Đại học Điện lực.

Lĩnh vực nghiên cứu: vật liệu điện tử, mạch điện tử, cảm biến, cảm biến sinh học, điện tử y sinh.



Tác giả Đinh Văn Quang tốt nghiệp đại học năm 2020 ngành điện tử y tế, nhận bằng Thạc sĩ ngành kỹ thuật điện tử năm 2022 tại Trường Đại học Điện lực.

Lĩnh vực nghiên cứu: ứng dụng trí tuệ nhân tạo trong chẩn đoán và điều trị y tế, điện tử y sinh.



Tác giả Bùi Văn Tấn tốt nghiệp đại học ngành hệ thống điện năm 2013 tại Đại học Bách khoa Hà Nội, nhận bằng Thạc sĩ ngành kỹ thuật điện tử năm 2022 tại Trường Đại học Điện lực; đã tham gia nghiên cứu khoa học theo hướng học thuật "ứng dụng mạng neuron network trong y tế" được 2 năm.

Lĩnh vực nghiên cứu: ứng dụng trí tuệ nhân tạo trong chẩn đoán và điều trị y tế, điện tử y sinh.