SO SÁNH THUẬT GIẢI LAN TRUYỀN NGƯỢC VÀ MÁY HỌC CỰC ĐỘ TRONG PHÂN TÍCH DỮ LIỆU Y KHOA

Huỳnh Trung Hiếu*

TÓM TẮT

Mạng neural nhân tạo là một trong những công cụ rất mạnh trong phân tích dữ liệu với một loạt các mô hình và các cải tiến được đề nghị. Do đó việc đánh giá, so sánh các thuật toán đóng vai trò hết sức quan trọng, giúp các nhà nghiên cứu có cái nhìn chính xác hơn và chọn cách tiếp cận thích hợp cho bài toán ứng dụng cụ thể. Trong bài báo này, tác giả trình bài một sự so sánh, đánh giá giữa thuật toán lan truyền ngược và thuật toán máy học cực độ đã được đề nghị gần đây trên các bài toán phân tích dữ liệu y khoa. Qua đó cung cấp cho người đọc cũng như các nhà nghiên cứu có cái nhìn bao quát hơn hiệu quả của các thuật toán huấn luyện mạng.

A COMPARISON OF BACKPROPAGATION ALGORITHM AND EXTREME LEARNING MACHINE IN MEDICAL DATA ANALYSIS

SUMMARY

Neural network is one of powerful tools in data analysis. Several models and improvements have been proposed. In this paper, the evaluation and comparison between the back-propagation and extreme learning machine algorithms on medical data analysis are presented. This plays an important role in choosing proper models and algorithms of neural networks for many different applications; especially for applications of medical data analysis.

1. GIỚI THIỆU

Phân tích dữ liệu y khoa đóng một vai trò hết sức quan trọng trong việc nâng cao hiệu quả điều trị và chăm sóc sức khỏe con người. Cùng với sự phát triển của nhiều ngành khác nhau, công nghệ thông tin đã và đang có những đóng góp rất tích cực trong lĩnh vực này. Một trong những công cụ được sử dụng phổ biến đó là máy học, cho phép tích hợp kiến thức chuyên gia vào các hệ thống nhằm giúp bác sĩ có thể chẩn đoán chính xác hơn và nhanh hơn.

Nhiều phương pháp tiếp cận máy học đã được đề nghị như các phương pháp thống kê, support vector machine (SVM) hoặc mạng neural,… Các phương pháp thống kê thường yêu cầu kiến thức trước về phân bố của dữ liệu, điều này không dễ được áp dụng cho nhiều bài toán. Các tiếp cận SVM thường gặp khó khăn trong việc chọn mô hình thích hợp. Đối với mạng neural, hiệu quả của nó đã được chứng minh qua nhiều ứng dụng thuộc rất nhiều lĩnh vực khác nhau.

Một vấn đề quan trọng trong mạng neural là chọn thuật toán huấn luyện mạng thích hợp. Trước kia, người ta thường sử dụng thuật toán giảm gradient. Tiếp cận này tồn tại nhiều vấn đề. Có nhiều cải tiến khác nhau đã được đề nghị để cải tiến các tiếp cận giảm gradient [1-5]. Nguyen và Widrow [1] đã đề nghị một phương pháp chọn các trọng số khởi động để tăng tốc độ hội tụ của lời giải. Bên cạnh gradient bậc nhất, những thuật giải lan truyền ngược dựa trên gradient bậc 2 cũng đã được nghiên cứu và phát triển [5]. Ngoài ra, cũng có rất nhiều phương pháp được đưa ra để khắc phục vấn đề overfitting trong huấn luyện mạng neural. Gần đây, G.-B Huang và các cộng sự đã đề nghị một thuật toán học khá hiệu quả là máy học cực độ (ELM). Nó có thể đạt độ chính xác cao với tốc độ học cực nhanh trong nhiều ứng dụng khác nhau [6, 7].

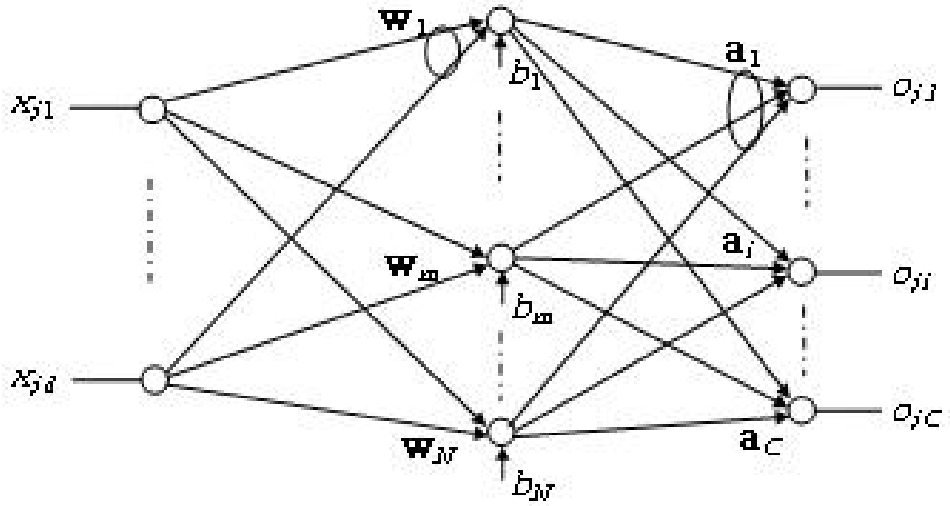
* TS. GV. Khoa công nghệ thông tin - trường Đại học Công nghiệp thành phố HCM

Trong bài báo này, tác giả trình bài sự so sánh giữa các thuật toán lan truyền ngược dựa trên giảm gradient và thuật toán máy học cực độ cho các ứng dụng phân tích dữ liệu Y khoa. Qua đó cung cấp một cái nhìn chính xác hơn về các tiếp cận cho ứng dụng mạng neural.

2. MẠNG NEURAL MỘT LỚP ẨN VÀ CÁC THUẬT TOÁN HUẤN LUYỆN

2.1. Mạng neural một lớp ẩn (SLFN)

Có nhiều kiến trúc mạng khác nhau đã và đang được nghiên cứu và phát triển. Tuy nhiên người ta đã chứng minh được rằng một mạng neural truyền thẳng với lớp ẩn đơn có thể tạo ra các biên phân loại với hình dạng bất kỳ nếu hàm tác động được chọn một cách thích hợp. Do đó, mạng một lớp ẩn đã và đang được ứng dụng phổ biến nhất. Kiến trúc tiêu biểu của mạng neural một lớp ẩn với *d* nút ở lớp nhập, *N* nút ở lớp ẩn và *C* nút ở lớp xuất có thể được mô tả như trong hình 1:



Hình 1. *Kiến trúc tiêu biểu của mạng neural một lớp ẩn (SLFN).*

Giả sử $\mathbf{w}_m = [w_{m1}, w_{m2}, ..., w_{md}]$ là vector trọng số của các kết nối từ lớp nhập đến nút ẩn thứ *m*, b_m là độ dịch của nó và $\mathbf{a}_i = [a_{i1}, a_{i2}, ..., a_{iN}]$ là vector trọng số của các kết nối từ lớp ẩn đến nút xuất thứ *i*. Thì vector ngõ xuất $\mathbf{o}_j$ tương ứng với vector nhập $\mathbf{x}_j$ được xác định bởi

$$\mathbf{o}_{ji} = \sum_{m=1}^{N} a_{im} f(\mathbf{w}_m \cdot \mathbf{x}_j + b_m),\ \mathbf{x} \in \mathbb{R}^d \qquad (1)$$

Trong đó $f(\cdot)$ là hàm tác động của các nút ẩn, $\mathbf{w}_m \cdot \mathbf{x} = \langle \mathbf{w}_m, \mathbf{x} \rangle$ là tích nội giữa 2 vector $\mathbf{w}_m$ và $\mathbf{x}$.

Cho tập mẫu $S = \{(\mathbf{x}_j, \mathbf{t}_j) \mid j = 1,\ldots,2\}$, mục đích chính của quá trình huấn luyện mạng là tìm ra các trọng số, bao gồm $\mathbf{w}$, $\mathbf{a}$ và b, để tối ưu một hàm mục tiêu nào đó. Thông thường, hàm mục tiêu được chọn là bậc 2 được định nghĩa như sau:

$$E = \sum_{j=1}^{n} \left(\mathbf{o}_j - \mathbf{t}_j \right)^2$$

$$= \sum_{j=1}^{n} \left(\sum_{m=1}^{N} \mathbf{a}_{im} f(\mathbf{w}_i \cdot \mathbf{x}_j + b_i) - \mathbf{t}_j \right)^2 \qquad (2)$$

2.2. Thuật toán lan truyền ngược

Lời giải cho (2) thường được tìm thông qua giảm gradient, trong đó các trọng số của mạng được xác định thông qua công thức lặp:

$$\mathbf{w}_{\mathrm{k}} = \mathbf{w}_{\mathrm{k}} - \mu \frac{\partial E}{\partial \mathbf{w}} \qquad (3)$$

với μ được gọi là hệ số tốc độ học (learning rate). Nó thường được sử dụng để tăng tốc độ hội tụ. Ngoài ra, thông số động lực học (momentum) cũng có thể được thêm vào nhằm tăng hiệu quả của quá trình tìm trọng số mạng.

Một trong những thuật toán phổ biến cho mạng neural truyền thẳng dựa trên sự giảm gradient là thuật toán lan truyền ngược (backpropagation). Ở đó gradient của hàm mục tiêu được tính và trọng số của mạng được hiệu chỉnh dựa trên sự lan truyền lỗi từ lớp xuất đến lớp nhập. Có nhiều cải tiến khác nhau được đưa ra bởi nhiều nhà nghiên cứu [1-5]. D. Nguyen và B. Widrow [1] đã đề nghị cách khởi động các giá trị trọng số để nâng cao tốc độ học. Bên cạnh gradient bậc nhất, những thuật giải lan truyền ngược dựa trên gradient bậc 2 cũng đã được nghiên cứu và phát triển [5]. Ngoài ra, cũng có rất nhiều phương pháp được đưa ra để khắc phục vấn đề overfitting trong huấn luyện mạng neural. Tuy nhiên đến thời điểm hiện nay phần lớn các tiếp cận dựa trên giảm gradient gặp phải các vấn đề sau:

- Có thể bị overtraining, từ đó dẫn đến kết quả không tốt.
- Có thể bị mắc kẹt tại những điểm tối ưu cục bộ, thay vì tối ưu toàn cục.
- Có thể hội tụ rất chậm nếu như hệ số tốc độ học nhỏ. Tuy nhiên, nếu hệ số tốc độ học lớn thì có thể dẫn đến sự không ổn định.
- Mặc dù có rất nhiều cải tiến cho thuật giải lan truyền ngược, tuy nhiên đến nay nó vẫn tốn nhiều thời gian để xác trọng số của mạng.

2.3. Máy học cực độ

Một trong những thuật toán huấn luyện hiệu quả được phát triển gần đây là máy học cực độ hay ELM (extreme learniing machine). Nó dựa trên ý tưởng là thay vì xác định tất cả các trọng số mạng bằng các quá trình lặp lại, trọng số lớp nhập và độ lệch có thể được chọn ngẫu nhiên và trọng số lớp xuất được xác định bằng các bước đơn. Rõ ràng một mạng với N nút ẩn có thể xấp xỉ N mẫu với lỗi bằng 0, nghĩa là tồn tại các trọng số $\mathbf{w}$, $\mathbf{a}$ và b sao cho

$$\mathbf{t}_{\mathrm{j}} = \sum_{m=1}^{N} \mathbf{a}_{\mathrm{im}} f(\mathbf{w}_i \cdot \mathbf{x}_j + b_i),\ j = 1, 2, ..., N \qquad (4)$$

Phương trình này có thể được viết lại như sau:

$$\mathbf{HA}=\mathbf{T}. \qquad (5)$$

Trong đó $\mathbf{H}$ còn được gọi là ma trận ngõ xuất lớp ẩn, $\mathbf{T}=[\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_n]^T$ và $\mathbf{A}=[\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_N]^T$. Trong [7], các tác giả đã chứng minh được rằng ma trận $\mathbf{H}$ là khả đảo nếu số mẫu trong tập huấn luyện bằng số nút ẩn và hàm tác động khả vi phân. Trong trường hợp số nút ẩn nhỏ hơn số mẫu huấn luyện thì ma trận trọng số xuất $\mathbf{A}$ sẽ được xác định bởi ma trận giả đảo của $\mathbf{H}$ với sự chọn lựa ngẫu nhiên của trọng số nhập và độ dịch. Các kết quả này đã được chứng minh trong [7]. Như vậy, thuật giải ELM có thể được tóm tắt như sau:

- Gán các giá trị ngẫu nhiên cho trọng số nhập và độ dịch các nút ẩn.
- Tính ma trận ngõ xuất lớp ẩn $\mathbf{H}$.
- Xác định trọng số xuất bằng cách sử dụng phương trình sau:

$$\mathbf{A}=\mathbf{H}^{\dagger}\mathbf{T} \qquad (6)$$

trong đó $\mathbf{H}^{\dagger}$ được gọi là ma trận giả đảo của $\mathbf{H}$. Như vậy, các trọng số của mạng có thể được xác định bởi những bước đơn giản và không cần sự tính toán bởi các bước lặp như các thuật toán giảm gradient. Nó có thể khắc phục những nhược điểm như chọn lựa hệ số tốc độ học, epochs, khởi động giá trị ban đầu .v.v. Đặc biệt

thuật toán này cho thời gian huấn luyện rất nhanh. So sánh về hiệu quả của thuật toán này và thuật toán lan truyền ngược trên các tập dữ liệu thực tiếp tục thảo luận trong phần tiếp theo.

3. KẾT QUẢ THỰC NGHIỆM

Trong phần này, tác giả trình bài các thực nghiệm trên bốn tập dữ liệu y khoa bao gồm chuẩn đoán bệnh tiểu đường (diabetes), chuẩn đoán bệnh ung thư máu (leukemia), chuẩn đoán bệng ung thư vú (breast cancer) và chuẩn đoán bệnh ung thư tuyến tiền liệt (prostate cancer). Mô tả của các tập dữ liệu này được chỉ ra trong bảng 1.

Bảng 1. *Mô tả của các tập dữ liệu*

Tập dữ liệu	**Số thuộc tính**	**Số lớp**	**Số mẫu**
Diabetes	8	2	768
Leukemia	7,129	2	72
Beast cancer	24,188	2	97
Prostate cancer	12,600	2	136

Tập dữ liệu diabetes [8] đã được sử dụng trong nghiên cứu dấu hiệu bệnh tiểu đường theo tiêu chí của tổ chức sức khỏe thế giới (WHO). Nó bao gồm 768 mẫu của các bệnh nhân. Mỗi mẫu có 8 thuộc tính nhập với các giá trị trong đoạn [0 1] được phân loại để xác định xem bệnh nhân đó có dấu hiệu bệnh tiểu đường hay không. 75% của tập dữ liệu được dùng cho huấn luyện và 25% còn lại được dùng cho đánh giá.

Tập dữ liệu leukemia bao gồm 38 mẫu tủy xương được dùng trong huấn luyện mạng và 34 mẫu được dùng để đánh giá kết quả. Số thuộc tính của tập dữ liệu này là 7,129. Chi tiết của tập dữ liệu này có thể tham khảo trong [9].

Tập dữ liệu breast cancer chứa 97 mẫu bệnh, trong đó 46 mẫu có dấu hiệu phát triển nhanh sau năm năm và 51 mẫu còn lại tương ứng với trường hợp mà bệnh nhân vẫn khỏe mạnh sau năm năm phát hiện bệnh. Mục tiêu của nghiên cứu trên dữ liệu này là dự đoán khả năng phát triển bệnh, từ đó có thể đưa ra các giải pháp trị liệu thích hợp. Trong thực nghiệm, 78 mẫu được dùng cho huấn luyện và 19 mẫu còn lại được dùng trong đánh giá kết quả. Chi tiết của tập dữ liệu này có thể tham khảo trong [10].

Trong tập dữ liệu prostate cancer [11], tập huấn luyện chứa các expression profiles chất lượng cao được trích ra từ 52 mẫu khối u tuyến tiền liệt và 50 mẫu bình thường. Mỗi mẫu chứa probes của khoảng 12600 genes và ESTs. Tập đánh giá có 34 mẫu, trong đó 9 mẫu là bình thường và 25 mẫu bệnh. Mục tiêu áp dụng trong tập dữ liệu này là phân biệt các mẫu bệnh từ các mẫu không bệnh.

Các thực nghiệm được hiện thực trên môi trường Matlab 7.0, hàm tác động là *sigmoid*. Số nút ẩn được kiểm tra và tăng từng bước bởi 2, và giá trị tương đối tối ưu được xác định dựa trên cross-validation.

Bảng 2. *Kết quả so sánh của thuật toán lan truyền ngược và máy học cực độ*

Tập dữ liệu	Thuật toán	Thời gian huấn luyện (s)	Độ chính xác (%)		Số nút ẩn
			Tập huấn luyện	Tập kiểm tra	
Diabetes	Lan truyền ngược	3.1130	81.80±1.93	75.25±3.17	4
	ELM	0.0109	78.60±1.19	77.53±2.80	20
Prostate	Lan truyền ngược	33.22	95.09±11.80	83.24±13.37	2
	ELM	0.1321	78.63±3.36	59.11±8.48	30
Leukemia	Lan truyền ngược	14.102`	98.80±9.96	88.50±14.27	2
	ELM	0.0230	91.35±5.10	67.70±11.10	20
Beast cancer	Lan truyền ngược	53.9381	97.80±3.90	61.47±10.95	2
	ELM	0.2501	84.97±4.01	61.37±12.48	30

Kết quả trung bình của 50 lần thử được chỉ ra trong bảng 2. Có thể thấy rằng, đối với các tập dữ liệu có số thuộc tính nhỏ như diabetes thì ELM cho kết quả tốt hơn thuật toán lan truyền ngược. Đối với các tập dữ liệu có số thuộc tính lớn như microarray thì thuật toán lan truyền ngược lại cho kết quả tốt hơn. Người đọc có thể thấy rằng thuật toán lan truyền ngược có thể đạt độ chính xác 88.50% và 83.24% đối với tập dữ liệu chuẩn đoán bệnh ung thư máu (leukemia) và ung thư tuyến tiền liệt, trong khi thuật toán ELM chỉ đạt độ chính xác 67.70% và 59.11%.

Xét về mặt thời gian huấn luyện, chúng ta có thể thấy rằng thuật toán ELM nhanh gấp hàng trăm đến hàng chục ngàn lần so với thuật toán lan truyền ngược. Kết quả này là do thuật toán ELM chỉ thực hiện những bước đơn, trong khi thuật toán lan truyền ngược phải thực hiện rất nhiều bước lặp để tìm các giá trị trọng số mạng. Tuy nhiên, thuật toán ELM thường yêu cầu số nút ẩn lớn hơn, điều này dẫn đến mạng có độ phức tạp cao hơn.

4. KẾT LUẬN

Mạng neural là một trong những công cụ khá mạnh trong phân tích dữ liệu y khoa. Một loạt kiến trúc mạng và các thuật toán đã được đề nghị. Bài báo này cung cấp một cái nhìn tương đối về tính hiệu quả các thuật toán huấn luyện cho mạng neural truyền thẳng một lớp ẩn.

Thuật toán ELM có thể đạt được tốc độ rất cao trong huấn luyện và có thể khắc phục một số vấn đề thường gặp trong thuật toán lan truyền ngược như chọn lựa các thông số learning rate, epochs, momentum, và overtraining. Tuy nhiên nó lại thường yêu cầu số nút ẩn lớn hơn so với thuật toán lan truyền ngược. Từ các kết quả thực nghiệm chúng ta cũng thấy rằng, thuật toán ELM cho kết quả khá tốt đối với các tập dữ liệu có số thuộc tính nhỏ. Đối với các tập dữ liệu có số thuộc tính lớn thì thuật toán lan truyền ngược lại cho kết quả tốt hơn.

TÀI LIỆU THAM KHẢO

[1] D. Nguyen and B. Widrow, Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights, *Int'l Joint Conf. Neural Networks*, Vol. 3 (San Diego, CA, 1990), pp. 21–26.

[2] Jim Y. F. Yam and Tommy W. S. Chow, Feedforward networks training speed enhancement by optimal initialization of the synaptic coefficients, *IEEE Trans. on Neural Networks* **12**(2) (2001) 430–434.

[3] Karayiannis and A. N. Venetsanopoulos, "Artificial neural networks: Learning algorithms, performance evaluation, and applications, *Kluver Academic*, Boston, MA, (1993).

[4] Y. LeCun, L. Bottou, G. B. Orr and K.-R. M¨uller, Efficient backprop, *Lecture Notes in Computer Science* **1524 (**1998) 9–50.

[5] Syed Muhammad Aqil Burney, Tahseen Ahmed Jilani and Cemal Ardil, A comparison of first and second order training algorithms for artificial neural networks, *International Journal of Computational Intelligence* **1** (2004) 218–224.

[6] G.-B. Huang, Q.-Y. Zhu and C.-K. Siew, Extreme learning machine: A new learning scheme for feedforward neural networks, *Proc. of Int'l Joint Conf. on Neural Networks*, (July 2004).

[7] G.-B. Huang, Q.-Y. Zhu and C.-K. Siew, Extreme learning machine: Theory and applications, *Neurocomputing* **70** (2006) 489–501.

[8] C. J. Merz and P. M. Murphy, UCI Repository of machine learning databases, *Dept. Of Inform. Comp. Sci., Univ. California*. Available: http://mlearn.ics.uci.edu/databases/

[9] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield and E. S. Lander, Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science* **286**(5439) (1999) 531–537.

[10] L. J. Van, T. Veer, H. Dai, M. J. V. De Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. V. Der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards and S. H. Friend, Gene expression profiling predicts clinical outcome of breast cancer, *Nature* **415** (2002) 530–536.

[11] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. von D'Amico, J. P. Richie, E. S. Lander, M. Loda, P. W. Kantoff, T. R. Golub, and W. R. Sellers, "Gene expression correlates of clinical prostate cancer behavior", *Cancer Cell*, vol. 1, (2002) 203-209.