|  |  |  |
| --- | --- | --- |
|  | BAN CƠ YẾU CHÍNH PHỦ  “HỌC VIỆN KỸ THUẬT MẬT MÃ” | Mẫu 2 |

BÁO CÁO CHUYÊN ĐỀ SỐ 2.3

“Nghiên cứu về mạng nơ-ron và học máy sử dụng mạng nơ-ron”

NHIỆM VỤ: “Nghiên cứu và ứng dụng nền tảng học sâu để xây dựng hệ thống phát hiện mã độc trực tuyến”.

Mã số: 06/2022/CB.

Cơ quan chủ trì: Học viện Kỹ thuật Mật mã

Chủ nhiệm: ThS. Lê Đức Thuận

Hà Nội - 2022

|  |  |  |
| --- | --- | --- |
|  | BAN CƠ YẾU CHÍNH PHỦ  “HỌC VIỆN KỸ THUẬT MẬT MÃ” |  |

BÁO CÁO CHUYÊN ĐỀ SỐ 2.3

“Nghiên cứu về mạng nơ-ron và học máy sử dụng mạng nơ-ron”

NHIỆM VỤ: “Nghiên cứu và ứng dụng nền tảng học sâu để xây dựng hệ thống phát hiện mã độc trực tuyến”.

Mã số: 06/2022/CB.

Cơ quan chủ trì: Học viện Kỹ thuật Mật mã

Chủ nhiệm: ThS. Lê Đức Thuận

|  |  |
| --- | --- |
| **Người thực hiện chuyên đề** | **Cơ quan chủ trì** |
| *(Họ tên và chữ ký)* | *(Họ tên và chữ ký)* |

Hà Nội - 2022

MỤC LỤC

[MỤC LỤC 3](#_Toc115870784)

[DANH MỤC HÌNH ẢNH 4](#_Toc115870785)

[DANH MỤC BẢNG 5](#_Toc115870786)

[1. GIỚI THIỆU VỀ MẠNG NƠRON NHÂN TẠO 6](#_Toc115870787)

[1.1. Khái niệm, mô hình và khả năng ứng dụng 6](#_Toc115870788)

[1.1.1. Khái niệm [5] 6](#_Toc115870789)

[1.1.2. Mô hình mạng nơron nhân tạo [6] 10](#_Toc115870790)

[1.1.3. Khả năng ứng dụng của mạng nơron nhân tạo 14](#_Toc115870791)

[1.2. Mạng nơron truyền thẳng nhiều lớp [1] 16](#_Toc115870792)

[1.2.1. Một số vấn đề cần chú ý khi sử dụng mạng MLP 18](#_Toc115870793)

[2. MỘT SỐ GIẢI THUẬT TỐI ƯU HÓA TRỌNG SỐ MẠNG NƠRON NHÂN TẠO 21](#_Toc115870794)

[2.1. Thuật toán học theo phương pháp lan truyền ngược sai số [3] 21](#_Toc115870795)

[2.1.1. Mô hình của thuật toán lan truyền ngược sai số 21](#_Toc115870796)

[2.1.2. Một số yếu tố ảnh hưởng đến quá trình học theo phương pháp lan truyền ngược 27](#_Toc115870797)

[2.2. Giải thuật di truyền [2] 29](#_Toc115870798)

[2.2.1. Các khái niệm cơ bản 31](#_Toc115870799)

[2.2.2. Mô hình giải thuật di truyền 32](#_Toc115870800)

[2.2.3. Áp dụng giải thuật Di truyền cho bài toán mạng nơron nhân tạo 32](#_Toc115870801)

[2.3. Kết hợp các giải thuật 33](#_Toc115870802)

[TÀI LIỆU THAM KHẢO 37](#_Toc115870803)

# DANH MỤC HÌNH ẢNH

[Hình 1.1 Cấu tạo của tế bào nơron sinh học 7](#_Toc115870772)

[Hình 1.2 Nơron nhân tạo 7](#_Toc115870773)

[Bảng 1.1 Các hàm kích hoạt thường được sử dụng 8](#_Toc115870774)

[Hình 1.3 Mạng nơron chỉ có 1 nút và có sự phản hồi 11](#_Toc115870775)

[Hình 1.4 Mạng nơron truyền thẳng 1 lớp (Single-layer feedforward network) 11](#_Toc115870776)

[Hình 1.5 Mạng nơron hồi quy 1 lớp 11](#_Toc115870777)

[Hình 1.6 Mạng MLP tổng quát 12](#_Toc115870778)

[Hình 1.7 Hàm sigmoid g(x) = 1/(1+e-x) 18](#_Toc115870779)

[Hình 2.1 Lan truyền tín hiệu trong quá trình học theo pp lan truyền ngược sai số 22](#_Toc115870780)

[Hình 2.2 Sai số E được xét là hàm của trọng số W 23](#_Toc115870781)

[Hình 2.3 Minh họa về ý nghĩa của quán tính trong thực tế 28](#_Toc115870782)

[Hình 2.4: Tham số của giải thuật lan truyền ngược sai số: 34](#_Toc115870783)

# DANH MỤC BẢNG

[Bảng 2.1 Giá trị đầu vào và ra của bài toán XOR 33](#_Toc115807930)

[Bảng 2.2 So sánh khả năng hội tụ của mạng khi sử dụng hai phương pháp học GA và BP với sai số dừng lặp khác nhau 36](#_Toc115807931)

# 1. GIỚI THIỆU VỀ MẠNG NƠRON NHÂN TẠO

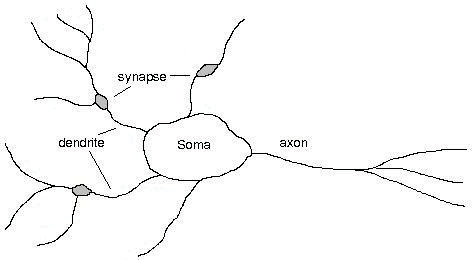
## 1.1. Khái niệm, mô hình và khả năng ứng dụng

### 1.1.1. Khái niệm [5]

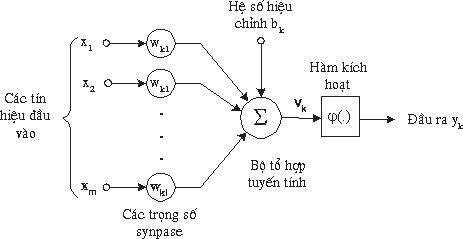
Theo các nhà nghiên cứu sinh học, hệ thống thần kinh của con người bao gồm khoảng 1011 tế bào thần kinh, thường gọi là các nơron. Mỗi tế bào nơron gồm ba phần:

* Thân nơron với nhân bên trong (gọi là soma), là nơi tiếp nhận hay phát ra các xung động thần kinh.
* Một hệ thống dạng cây các dây thần kinh vào (gọi là dendrite) để đưa tín hiệu tới nhân nơron. Các dây thần kinh vào tạo thành một lưới dày đặc xung quanh thân nơron, chiếm diện tích khoảng 0,25 mm2
* Đầu dây thần kinh ra (gọi là sợi trục axon) phân nhánh dạng hình cây, có thể dài từ một cm đến hàng mét. Chúng nối với các dây thần kinh vào hoặc trực tiếp với nhân tế bào của các nơron khác thông qua các khớp nối (gọi là synapse). Có hai loại khớp nối, khớp nối kích thích (excitatory) sẽ cho tín hiệu qua nó để tới nơron còn khớp nối ức chế (inhibitory) có tác dụng làm cản tín hiệu tới nơron. Người ta ước tính mỗi nơron trong bộ não của con người có khoảng 104 khớp nối.

Chức năng cơ bản của các tế bào nơron là liên kết với nhau để tạo nên hệ thống thần kinh điều khiển hoạt động của cơ thể sống. Các tế bào nơron truyền tín hiệu cho nhau thông qua các dây thần kinh vào và ra, các tín hiệu đó có dạng xung điện và được tạo ra từ các quá trình phản ứng hoá học phức tạp. Tại nhân tế bào, khi điện thế của tín hiệu vào đạt tới một ngưỡng nào đó thì nó sẽ tạo ra một xung điện dẫn tới trục dây thần kinh ra. Xung này truyền theo trục ra tới các nhánh rẽ và tiếp tục truyền tới các nơron khác.

Hình 1.1 Cấu tạo của tế bào nơron sinh học

Mạng nơron nhân tạo, Artificial Neural Network (ANN) gọi tắt là mạng nơron, neural network, là một mô hình xử lý thông tin phỏng theo cách thức xử lý thông tin của các hệ nơron sinh học. Nó được tạo nên từ một số lượng lớn các phần tử (gọi là phần tử xử lý hay nơron) kết nối với nhau thông qua các liên kết (gọi là trọng số liên kết) làm việc như một thể thống nhất để giải quyết một vấn đề cụ thể nào đó. Một nơron là một đơn vị xử lý thông tin và là thành phần cơ bản của một mạng nơron.

Cấu trúc của một nơron được mô tả trên hình 1.2.

Hình 1.2 Nơron nhân tạo

Các thành phần cơ bản của một nơron nhân tạo bao gồm:

* Tập các đầu vào: Là các tín hiệu vào (input signals) của nơron, các tín hiệu này thường được đưa vào dưới dạng một vec-tơ m chiều.
* Tập các liên kết: Mỗi liên kết được thể hiện bởi một trọng số (gọi là trọng số liên kết – Synaptic weight). Trọng số liên kết giữa tín hiệu vào thứ j với nơron k thường được kí hiệu là wjk. Thông thường, các trọng số này được khởi tạo một cách ngẫu nhiên ở thời điểm khởi tạo mạng và được cập nhật liên tục trong quá trình học mạng.
* Bộ tổng (Summing function): Thường dùng để tính tổng của tích các đầu vào với trọng số liên kết của nó.
* Ngưỡng (còn gọi là một độ lệch - bias): Ngưỡng này thường được đưa vào như một thành phần của hàm truyền.
* Hàm truyền (Transfer function) – còn gọi là Hàm kích hoạt (Activation function): Hàm này được dùng để giới hạn phạm vi đầu ra của mỗi nơron. Nó nhận đầu vào là kết quả của hàm tổng và ngưỡng đã cho. Thông thường, phạm vi đầu ra của mỗi nơron được giới hạn trong đoạn [0,1] hoặc [-1, 1]. Các hàm truyền rất đa dạng được liệt kê trong bảng 1.1, có thể là các hàm tuyến tính hoặc phi tuyến. Việc lựa chọn hàm truyền nào là tuỳ thuộc vào từng bài toán và kinh nghiệm của người thiết kế mạng.

Bảng 1.1 Các hàm kích hoạt thường được sử dụng

|  |  |  |
| --- | --- | --- |
| **ST T** | **Tên hàm** | **Phương trình** |
| 1 | Identity |  |
| 2 | Binary step |  |
| 3 | Logistic |  |
| 4 | TanH |  |
| 5 | ArcTan |  |
| 6 | Softsign |  |
| 7 | Rectified linear unit (ReLU) |  |
| 8 | Leaky rectified linear unit  (Leaky ReLU) |  |
| 9 | Parameteric rectified linear  unit (PReLU) |  |
| 10 | Randomized leaky rectified linear unit  (RReLU) |  |
| 11 | Exponential linear unit (ELU) |  |
| 12 | Scaled exponential  linear unit (SELU) |  |
| 13 | S-shaped rectified linear activation unit  (SReLU) |  |
| 14 | Adaptive  piecewise linear (APL) |  |
| 15 | SoftPlus |  |
| 16 | Bent identity |  |
| 17 | SoftExponential |  |
| 18 | Sinusoid |  |
| 19 | Sinc |  |
| 20 | Gaussian |  |

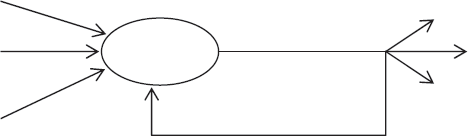
Đầu ra: Là tín hiệu đầu ra của một nơron, với mỗi nơron sẽ có tối đa là một đầu ra. Như vậy tương tự như nơron sinh học, nơron nhân tạo cũng nhận các tín hiệu đầu vào, xử lý (nhân các tín hiệu này với trọng số liên kết, tính tổng các tích thu được rồi gửi kết quả tới hàm truyền), và cho một tín hiệu đầu ra (là kết quả của hàm truyền).

### 1.1.2. Mô hình mạng nơron nhân tạo [6]

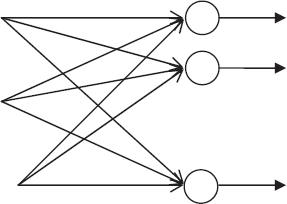
Mặc dù mỗi nơron đơn lẻ có thể thực hiện những chức năng xử lý thông tin nhất định, sức mạnh của tính toán nơron chủ yếu có được nhờ sự kết hợp các nơron trong một kiến trúc thống nhất. Một mạng nơron là một mô hình tính toán được xác định qua các tham số: kiểu nơron (như là các nút nếu ta coi cả mạng nơron là một đồ thị), kiến trúc kết nối (sự tổ chức kết nối giữa các nơron) và thuật toán học (thuật toán dùng để học cho mạng).

Về bản chất một mạng nơron có chức năng như là một hàm ánh xạ F: X → Y, trong đó X là không gian trạng thái đầu vào (input state space) và Y là không gian trạng thái đầu ra (output state space) của mạng. Các mạng chỉ đơn giản là làm nhiệm vụ ánh xạ các vec- tơ đầu vào x E X sang các vec-tơ đầu ra y E Y thông qua “bộ lọc” (filter) các trọng số. Tức là y = F(x) = s(W, x), trong đó W là ma trận trọng số liên kết. Hoạt động của mạng thường là các tính toán số thực trên các ma trận.

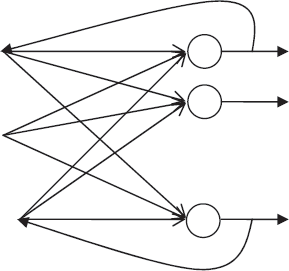
Trong bộ não của con người, các tế bào nơron liên kết với nhau chằng chịt và tạo nên một mạng lưới vô cùng phức tạp, tuy nhiên mạng nơron nhân tạo được chia thành các loại chính sau:

Hình 1.3 Mạng nơron chỉ có 1 nút và có sự phản hồi

Mạng nơron truyền thẳng một lớp (perceptron) là loại mạng chỉ có lớp nơron đầu vào và một lớp nơron đầu ra (thực chất lớp nơron đầu vào không có vai trò xử lý, do đó ta nói mạng chỉ có một lớp). Loại mạng này còn được gọi là mạng perceptron một lớp. Mỗi nơron đầu ra có thể nhận tín hiệu từ các đầu vào x1, x2, …, xm để tạo ra tín hiệu đầu ra tương ứng.

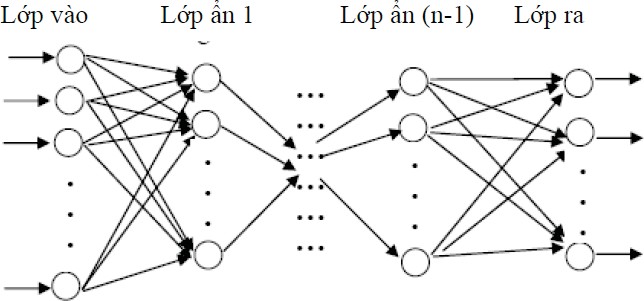


Hình 1.4 Mạng nơron truyền thẳng 1 lớp (Single-layer feedforward network)

Mạng có phản hồi (feedback network) là mạng mà đầu ra của một nơron có thể trở thành đầu vào của nơron trên cùng một lớp hoặc của lớp trước đó. Mạng feedback có chu trình khép kín gọi là mạng quy hồi (recurrent network).

Hình 1.5 Mạng nơron hồi quy 1 lớp

Mô hình mạng truyền thẳng nhiều lớp: Mô hình mạng nơron được sử dụng rộng rãi nhất là mô hình mạng nhiều tầng truyền thẳng (MLP: Multi Layer Perceptron). Một mạng MLP tổng quát là mạng có n (n ≥ 2) lớp (thông thường lớp đầu vào không được tính đến): trong đó gồm một lớp đầu ra (lớp thứ n) và (n-1) lớp ẩn.



Hình 1.6 Mạng MLP tổng quát

Một số kết quả đã được chứng minh với mạng MLP:

* Bất kì một hàm Boolean nào cũng có thể biểu diễn được bởi một mạng MLP 2
* lớp trong đó các nơron sử dụng hàm truyền sigmoid.
* Tất cả các hàm liên tục đều có thể xấp xỉ bởi một mạng MLP 2 lớp sử dụng hàm truyền sigmoid cho các nơron lớp ẩn và hàm truyền tuyến tính cho các nơron lớp ra với sai số nhỏ tùy ý.
* Mọi hàm bất kỳ đều có thể xấp xỉ bởi một mạng MLP 3 lớp sử dụng hàm truyền
* sigmoid cho các nơron lớp ẩn và hàm truyền tuyến tính cho các nơron lớp ra.

Mô hình mạng kết hợp : Trong thực tế, người ta thường kết hợp nhiều loại mạng để giải quyết các bài toán. Có thể tham khảo mô hình ANN hệ thống chơi cờ vây AlphaGo của Google bao gồm kết hợp 3 mạng ANN. Mạng 1 SL có cấu trúc xoắn bao gồm 13 lớp ẩn với tập dữ liệu học có giám sát là hàng triệu nước cờ đã có thu thập từ các ván cờ các kỳ thủ đã chơi. Mạng 2 RL có cấu trúc giống mạng SL nhưng được tối ưu hóa bộ trọng số qua bộ huấn luyện là các ván cờ tự chơi với nhau giữa các máy. Mạng 3 Value dự đoán kết quả trò chơi dựa trên RL, giá trị của vị trí p được định nghĩa như 1 kỳ vọng phân phối các kết quả phát từ vị trí p đến cuối ván cờ. Cuối cùng, AlphaGo kết hợp mạng trên với thuật toán Monte Carlo Tree Search để đưa ra đánh giá vị trí cuối cùng cho nước đi tiếp theo.

**Quá trình học của mạng nơron**

Học là quá trình cập nhật trọng số sao cho giá trị hàm lỗi là nhỏ nhất. Một mạng nơron được huấn luyện sao cho với một tập các vec-tơ đầu vào X, mạng có khả năng tạo ra tập các vec-tơ đầu ra mong muốn Y của nó. Tập X được sử dụng cho huấn luyện mạng được gọi là tập huấn luyện (training set). Các phần tử x thuộc X được gọi là các mẫu huấn luyện (training example). Quá trình huấn luyện bản chất là sự thay đổi các trọng số liên kết của mạng. Trong quá trình này, các trọng số của mạng sẽ hội tụ dần tới các giá trị sao cho với mỗi vec-tơ đầu vào x từ tập huấn luyện, mạng sẽ cho ra vec-tơ đầu ra y như mong muốn.

Có ba phương pháp học phổ biến là học có giám sát (*supervised learning*), học không giám sát (*unsupervised learning*) và học tăng cường (*Reinforcement learning*).

**Học có giám sát trong các mạng nơron**

Học có giám sát có thể được xem như việc xấp xỉ một ánh xạ: X→ Y, trong đó X là tập các vấn đề và Y là tập các lời giải tương ứng cho vấn đề đó. Các mẫu (x, y) với x = (x1, x2, ..., xn) E X, y = (yl, y2, ..., ym) E Y được cho trước. Học có giám sát trong các mạng nơron thường được thực hiện theo các bước sau:

* Bước 1: Xây dựng cấu trúc thích hợp cho mạng nơron, chẳng hạn có n nơron vào, m nơron đầu ra, và khởi tạo các trọng số liên kết của mạng.
* Bước 2: Đưa một vec-tơ x trong tập mẫu huấn luyện X vào mạng
* Bước 3: Tính vec-tơ đầu ra z của mạng
* Bước 4: So sánh vec-tơ đầu ra mong muốn t (là kết quả được cho trong tập huấn luyện) với vec-tơ đầu ra z do mạng tạo ra; nếu có thể thì đánh giá lỗi.
* Bước 5: Hiệu chỉnh các trọng số liên kết theo một cách nào đó sao cho ở lần tiếp theo khi đưa vec-tơ x vào mạng, vec-tơ đầu ra z sẽ giống với t hơn.
* Bước 6: Nếu cần, lặp lại các bước từ 2 đến 5 cho tới khi mạng đạt tới trạng thái hội tụ. Việc đánh giá lỗi có thể thực hiện theo nhiều cách, cách dùng nhiều nhất là sử dụng lỗi tức thời: Err = (z - t), hoặc Err = |z - t| ; lỗi trung bình bình phương (MSE: mean- square error): Err = (z- t)2/2.

Có hai loại lỗi trong đánh giá một mạng nơron. Thứ nhất, gọi là lỗi rõ ràng (apparent error), đánh giá khả năng xấp xỉ các mẫu huấn luyện của một mạng đã được huấn luyện. Thứ hai, gọi là lỗi kiểm tra (test error), đánh giá khả năng tổng quát hóa của một mạng đã được huấn luyện, tức khả năng phản ứng với các vec-tơ đầu vào mới. Để đánh giá lỗi kiểm tra chúng ta phải biết đầu ra mong muốn cho các mẫu kiểm tra.

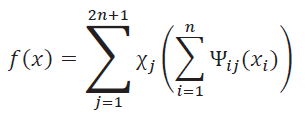
Thuật toán tổng quát ở trên cho học có giám sát trong các mạng nơron có nhiều cài đặt khác nhau, sự khác nhau chủ yếu là cách các trọng số liên kết được thay đổi trong suốt thời gian học. Trong đó tiêu biểu nhất là thuật toán lan truyền ngược sai số.

### 1.1.3. Khả năng ứng dụng của mạng nơron nhân tạo

Đặc trưng của mạng nơron nhân tạo là khả năng học. Nó có thể gần đúng mối quan hệ tương quan phức tạp giữa các yếu tố đầu vào và đầu ra của các quá trình cần nghiên cứu và khi đã học được thì việc kiểm tra độc lập thường cho kết quả tốt. Sau khi đã học xong, mạng nơron nhân tạo có thể tính toán kết quả đầu ra tương ứng với bộ số liệu đầu vào mới.

Về mặt cấu trúc, mạng nơron nhân tạo là một hệ thống gồm nhiều phần tử xử lý đơn giản cùng hoạt động song song. Tính năng này của ANN cho phép nó có thể được áp dụng để giải các bài toán lớn.

Về khía cạnh toán học, theo định lý Kolmogorov, một hàm liên tục bất kỳ f(x1, x2,…, xn ) xác định trên khoảng In ( với I =[0,1]) có thể được biểu diễn dưới dạng [4]:



trong đó: χj , Ψij là các hàm liên tục một biến. Ψij là hàm đơn điệu, không phụ thuộc vào hàm f. Mặt khác, mô hình mạng nơron nhân tạo cho phép liên kết có trọng số các phần tử phi tuyến (các nơron đơn lẻ) tạo nên dạng hàm tổng hợp từ các hàm thành phần. Do vậy, sau một quá trình điều chỉnh sự liên kết cho phù hợp (quá trình học), các phần tử phi tuyến đó sẽ tạo nên một hàm phi tuyến phức tạp có khả năng xấp xỉ hàm biểu diễn quá trình cần nghiên cứu. Kết quả là đầu ra của nó sẽ tương tự với kết quả đầu ra của tập dữ liệu dùng để luyện mạng. Khi đó ta nói mạng nơron nhân tạo đã học được mối quan hệ tương quan đầu vào - đầu ra của quá trình và lưu lại mối quan hệ tương quan này thông qua bộ trọng số liên kết giữa các nơron. Do đó, mạng nơron nhân tạo có thể tính toán trên bộ số liệu đầu vào mới để đưa ra kết quả đầu ra tương ứng.

Với những đặc điểm đó, mạng nơron nhân tạo đã được sử dụng để giải quyết nhiều bài toán thuộc nhiều lĩnh vực của các ngành khác nhau. Các nhóm ứng dụng mà mạng nơron nhân tạo đã được áp dụng rất có hiệu quả là:

* **Bài toán phân lớp**: Loại bài toán này đòi hỏi giải quyết vấn đề phân loại các đối tượng quan sát được thành các nhóm dựa trên các đặc điểm của các nhóm đối tượng đó. Đây là dạng bài toán cơ sở của rất nhiều bài toán trong thực tế: nhận dạng chữ viết, tiếng nói, phân loại gen, phân loại chất lượng sản phẩm, v.v.
* **Bài toán dự báo**: Mạng nơron nhân tạo đã được ứng dụng thành công trong việc xây dựng các mô hình dự báo sử dụng tập dữ liệu trong quá khứ để dự đoán số liệu trong tương lai. Đây là nhóm bài toán khó và rất quan trọng trong nhiều ngành khoa học.
* **Bài toán điều khiển và tối ưu hoá**: Nhờ khả năng học và xấp xỉ hàm mà mạng nơron nhân tạo đã được sử dụng trong nhiều hệ thống điều khiển tự động cũng như góp phần giải quyết những bài toán tối ưu trong thực tế.

Tóm lại, mạng nơron nhân tạo được xem như là một cách tiếp cận đầy tiềm năng để giải quyết các bài toán có tính phi tuyến, phức tạp và đặc biệt là trong tình huống mối quan hệ bản chất vật lý của quá trình cần nghiên cứu không dễ thiết lập tường minh.

## 1.2. Mạng nơron truyền thẳng nhiều lớp [1]

Mạng perceptron một lớp do F.Rosenblatt đề xuất năm 1960 là mạng truyền thẳng chỉ một lớp vào và một lớp ra không có lớp ẩn. Trên mỗi lớp này có thể có một hoặc nhiều nơron. Rosenblatt đã chứng minh rằng quá trình học của mạng Perceptron sẽ hội tụ tới bộ trọng số W, biểu diễn đúng các mẫu học với điều kiện là các mẫu này biểu thị các điểm rời rạc của một hàm khả tách tuyến tính nào đó (f: Rn →R được gọi là khả tách tuyến tính nếu các tập {F -1(xk)}, với xk thuộc miền trị của f, có thể tách được với nhau bởi các siêu phẳng trong không gian Rn).

Năm 1969, Minsky và Papert đã chứng minh một cách chặt chẽ rằng lớp hàm thể hiện sự phụ thuộc giữa đầu vào và đầu ra có thể học bởi mạng Perceptron một lớp là lớp hàm khả tách tuyến tính. Khả tách tuyến tính là trường hợp tồn tại một mặt siêu phẳng để phân cách tất cả các đối tượng của một lớp này với một lớp khác, ví dụ một mặt phẳng sẽ phân chia không gian ba chiều thành hai vùng riêng biệt. Mở rộng ra, nếu có n đầu vào, n>2 thì công thức In wij xj = 0i tạo nên một siêu phẳng có n-1 chiều trong không gian n chiều, nó chia không gian đó thành hai nửa. Trong nhiều bài toán thực tế đòi hỏi chia các vùng của các điểm trong một siêu không gian thành các lớp riêng biệt. Loại bài toán này gọi là bài toán phân lớp. Bài toán phân lớp có thể giải quyết bằng cách tìm các tham số thích hợp cho một siêu phẳng để nó có thể chia không gian n chiều thành các vùng riêng biệt.

j=1

Với tính chất của như đã nêu trên, mạng perceptron một lớp có thể mô tả các hàm logic như AND, OR và NOT. Tuy nhiên nó không thể hiện được hàm XOR. Như vậy chứng tỏ mô hình perceptron một lớp không thể giải quyết bài toán này. Vấn đề này sẽ được giải quyết bằng mô hình mạng nơron perceptron nhiều lớp (Multi Layer Perceptron - MLP)

Mạng perceptron nhiều lớp (Multilayer Perceptron – MLP) còn được gọi là mạng truyền thẳng nhiều lớp là sự mở rộng của mô hình mạng perceptron với sự bổ sung thêm những lớp ẩn và các nơron trong các lớp ẩn này có hàm truyền (hàm kích hoạt) dạng phi tuyến. Mạng MLP có một lớp ẩn là mạng nơron nhân tạo được sử dụng phổ biến nhất, nó có thể xấp xỉ các hàm liên tục được định nghĩa trên một miền có giới hạn cũng như những hàm là tập hợp hữu hạn của các điểm rời rạc.

Cấu trúc của một mạng MLP tổng quát có thể mô tả như sau:

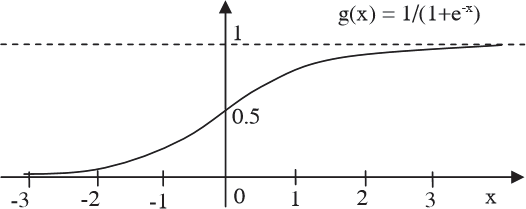
* Đầu vào là các vec-tơ (x1, x2, ..., xp) trong không gian p chiều, đầu ra là các vec- tơ (y1, y2, ..., yq) trong không gian q chiều. Đối với các bài toán phân loại, p chính là kích thước của mẫu đầu vào, q chính là số lớp cần phân loại.
* Mỗi nơron thuộc lớp sau liên kết với tất cả các nơron thuộc lớp liền trước nó.
* Đầu ra của nơron lớp trước là đầu vào của nơron thuộc lớp liền sau nó.

Hoạt động của mạng MLP như sau: tại lớp đầu vào các nơron nhận tín hiệu vào xử lý (tính tổng trọng số, gửi tới hàm truyền) rồi cho ra kết quả (là kết quả của hàm truyền); kết quả này sẽ được truyền tới các nơron thuộc lớp ẩn thứ nhất; các nơron tại đây tiếp nhận như là tín hiệu đầu vào, xử lý và gửi kết quả đến lớp ẩn thứ 2 ; … ; quá trình tiếp tục cho đến khi các nơron thuộc lớp ra cho kết quả.

### 1.2.1. Một số vấn đề cần chú ý khi sử dụng mạng MLP

Mạng nơron perceptron nhiều lớp là loại mạng nơron được sử dụng trong nhiều ứng dụng thực tế. Tuy nhiên, để mạng có thể đưa ra kết quả tốt, chúng ta cần quan tâm đến một số vấn đề có ảnh hưởng khá quan trọng đến hiệu quả làm việc của nó bao gồm: vấn đề chuẩn hoá số liệu đầu vào, vấn đề học chưa đủ và học quá của mạng, vấn đề lựa chọn một cấu trúc mạng phù hợp với bài toán.

#### 1.2.1.1. Vấn đề chuẩn hoá số liệu đầu vào

Mạng MLP thường sử dụng hàm truyền là hàm sigmoid có dạng như sau:

Hình 1.7 Hàm sigmoid g(x) = 1/(1+e-x)

Với dạng hàm này, giá trị ở đầu ra của mỗi nơron nằm trong phạm vi khoảng (0,1) và nó đạt các giá trị bão hoà (xấp xỉ 0 hay 1) khi Ix I lớn. Do đó, khi đầu vào của mạng có giá trị tuyệt đối lớn thì ta cần chuẩn hoá nó về khoảng có giá trị nhỏ, nếu không thì các nơron tại các lớp ẩn ngay ban đầu đã có thể đạt giá trị bão hoà và quá trình học của mạng không đạt kết quả mong muốn. Với dạng hàm như trên thì giá trị đầu vào của mạng thường được chuẩn hoá về khoảng thuộc đoạn [-3, 3]. Mặt khác, do tín hiệu đầu ra của nơron nằm trong khoảng giá trị (0,1) nên các giá trị đầu ra thực tế trong các mẫu học cũng cần chuẩn hoá về khoảng giá trị này để có thể dùng cho quá trình luyện mạng. Do vậy trong quá trình tính toán, để có các giá trị thực tế ở đầu ra của mạng chúng ta cần phải chuyển các giá trị trong khoảng (0,1) về miền các giá trị thực tế.

#### 1.2.1.2. Vấn đề học chưa đủ và học quá thuộc của mạng

Vấn đề mấu chốt khi xây dựng một mạng nơron nhân tạo là làm thế nào mạng có khả năng tổng quát hoá cao để đưa ra kết quả tốt cả với những trường hợp đầu vào của mạng không nằm trong tập mẫu đã dùng để luyện mạng. Giống như các mô hình hồi quy phi tuyến khác, đối với mạng nơron nhân tạo ta cũng phải giải quyết hai vấn đề là ANN học chưa đủ (underfitting) và học quá (overfitting). Khi mạng có cấu trúc (số nút ẩn và liên kết) cũng như số lần học chưa đủ so với nhu cầu của bài toán thì sẽ dẫn tới tình trạng mạng không đủ khả năng mô tả gần đúng mối quan hệ tương quan giữa đầu vào và đầu ra của quá trình cần dự báo và dẫn tới học chưa đủ. Trái lại, nếu mạng quá phức tạp (quá nhiều nút ẩn và quá nhiều tham số) và được học “quá khít” đối với các mẫu dùng để luyện mạng thì có thể dẫn tới tình trạng mạng học cả thành phần nhiễu lẫn trong các mẫu đó, đây là tình trạng “học quá thuộc” của mạng. Vấn đề nêu trên có thể làm cho nhiều loại mạng nơron, đặc biệt là mạng MLP có thể có những trường hợp cho kết quả dự đoán rất sai lệch với thực tế.

**Một số giải pháp cho vấn đề học quá của mạng:**

* Sử dụng tập số liệu có tính đại diện tốt để luyện mạng: Đây được xem là một cách khá tốt để tránh hiện tượng overfitting. Khi tập mẫu dùng để luyện mạng thể hiện được nhiều trạng thái có thể xẩy ra của quá trình cần nghiên cứu thì sau khi học mạng sẽ có khả năng tổng quát hoá tương đối tốt từ tập dữ liệu đó và sẽ không chịu ảnh hưởng nhiều của hiện tượng overfitting. Ngoài ra một số biện pháp dưới đây cũng có thể góp phần quan trọng giúp khắc phục hiện tượng overfitting của mạng.
* Lựa chọn cấu trúc mô hình phù hợp: Việc lựa chọn mô hình của mạng (số lớp ẩn, số nơron trên mỗi lớp ẩn) có ảnh hưởng quan trọng đến hiện tượng học chưa đủ (underfitting) và học quá (overfitting) của mạng. Các mạng có độ phức tạp hơn tuy nó có thể học khá chính xác các mẫu được sử dụng nhưng chính điều này lại làm cho nó học quá nhiều cả thành phần nhiễu nên khả năng tổng quát hoá giảm và dẫn tới hiện tượng học quá (overfitting).
* Dừng học đúng lúc: giải pháp dừng học đúng lúc để tránh hiện tượng học quá của mạng như sau:
* Tập mẫu được chia làm hai phần: một phần dùng để luyện mạng và phần còn lại để kiểm thử.
* Sử dụng các giá trị khởi tạo nhỏ.
* Sử dụng hằng số tốc độ học có giá trị thấp.
* Tính toán sự thay đổi lỗi kiểm thử trong quá trình luyện mạng.
* Dừng học khi thấy lỗi kiểm thử bắt đầu tăng.

#### 1.2.1.3. Lựa chọn kích thước mạng

Các công trình dựa trên định lý của Kolmogorov dự kiến rằng toàn bộ các ánh xạ liên tục từ [0,1]p đến [0,1]n đều có thể được xấp xỉ bằng một mạng perceptron ba lớp có lớp vào gồm p nơron, lớp ra gồm n nơron và lớp ẩn gồm (2p+1) nơron. Tuy nhiên không thể chỉ ra được chính xác số lượng nơron tối ưu trong mạng, tính chất của các nơron, tức là dạng phi tuyến cụ thể thực hiện phép xấp xỉ này. Một số công trình nghiên cứu về chủ đề này cho rằng số nơron tối ưu ở lớp ẩn thường nhỏ hơn (2p+1). Ngoài ra cũng cần phải nói cơ sở dữ liệu học phải có kích thước phù hợp với kiến trúc mạng. Theo Vapnik và Chervonenkis, cơ sở dữ liệu học phải có số mẫu thoả mãn: N ≈ 10.Nw, ở đó Nw là số trọng số của mạng. Gọi số nơron thuộc lớp ẩn là L, số nơron ở lớp vào là p thì trọng số của các kết nối giữa lớp vào và lớp ẩn thứ nhất (kể cả ngưỡng) là: D=(p+1).L. Theo một số kết quả nghiên cứu, số mẫu của cơ sở dữ liệu học cần phải thoả mãn: N ≈ 4.D. Khi số lượng mẫu của cơ sở dữ liệu học chưa đạt đến giới hạn cần thiết thì ta nên làm giảm số lượng các kết nối để tránh hiện tượng học thuộc lòng.

# 2. MỘT SỐ GIẢI THUẬT TỐI ƯU HÓA TRỌNG SỐ MẠNG NƠRON NHÂN TẠO

## 2.1. Thuật toán học theo phương pháp lan truyền ngược sai số [3]

Thuật toán học theo phương pháp lan truyền ngược sai số là một trong số những kết quả nghiên cứu quan trọng nhất đối với sự phát triển của mạng nơron nhân tạo. Thuật toán này được áp dụng cho mạng truyền thẳng nhiều lớp trong đó các nơron có thể sử dụng các hàm truyền là các hàm liên tục có các dạng khác nhau.

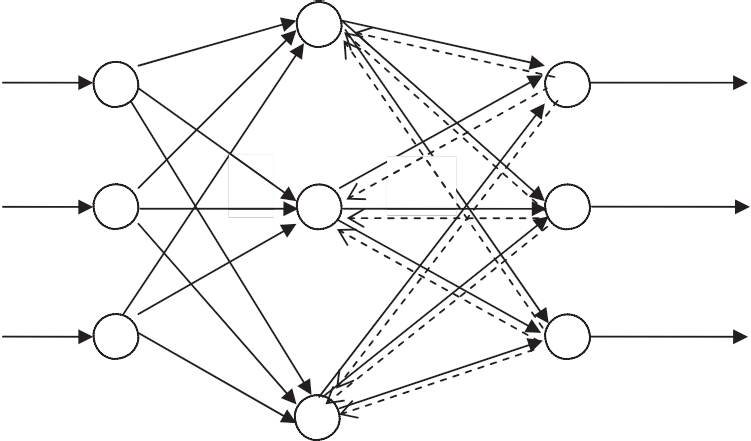
### 2.1.1. Mô hình của thuật toán lan truyền ngược sai số

Thuật toán sử dụng một tập các mẫu gồm các cặp đầu vào - đầu ra để luyện mạng. Với mỗi cặp đầu vào - đầu ra (x(k), d(k)) thuật toán lan truyền ngược sai số thực hiện hai giai đoạn sau:

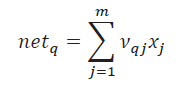
* Giai đoạn thứ nhất, mẫu đầu vào x(k) được truyền từ lớp vào tới lớp ra, và ta có kết quả đầu ra tính toán được là y(k).
* Giai đoạn tiếp theo, tín hiệu lỗi được tính toán từ sự khác nhau giữa đầu ra quan sát được d(k) với đầu ra tính toán y(k) sẽ được lan truyền ngược lại từ lớp ra đến các lớp trước để điều chỉnh các trọng số của mạng.

Để làm ví dụ ta xét mạng truyền thẳng có một lớp ẩn dưới đây, đối với các mạng có kích thước lớn hơn thì thao tác cũng tương tự.

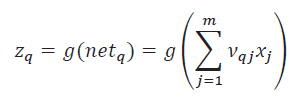
Mạng nơron được xét có m nơron ở lớp vào, l nơron trong lớp ẩn và n nơron ở lớp ra. Đường kẻ liền thể hiện luồng tín hiệu được truyền từ đầu vào tới đầu ra còn các đường kẻ nét đứt thể hiện luồng tín hiệu lỗi được truyền ngược trở lại từ đầu ra.

Hình 2.1 Lan truyền tín hiệu trong quá trình học theo pp lan truyền ngược sai số

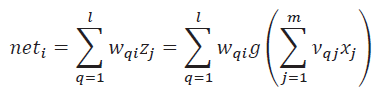
Chúng ta xét một cặp đầu vào - đầu ra để luyện mạng (x, d), để đơn giản chúng ta bỏ ký hiệu mũ k thể hiện số thứ tự của cặp mẫu này trong bộ mẫu dùng để luyện mạng. Khi đưa vào đầu vào x, nơron thứ q trong lớp ẩn sẽ nhận tín hiệu vào của mạng là:



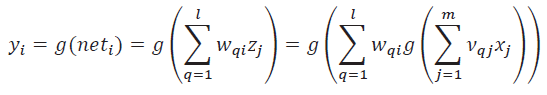
Nơron *q* ở lớp ẩn sẽ tính toán và tạo kết quả ở đầu ra của nó là:



Do đó tín hiệu vào của nơron thứ i trên lớp ra sẽ là:

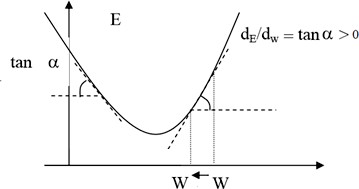


Và cuối cùng, đầu ra của nơron i trên lớp ra sẽ là :



Công thức trên cho biết quá trình lan truyền tín hiệu từ đầu vào qua lớp ẩn tới đầu ra. Tiếp theo chúng ta xét tín hiệu lỗi được lan truyền ngược lại từ lớp ra. Trước hết, đối với mỗi cặp giá trị vào – ra chúng ta xây dựng một hàm giá như sau :

Như vậy với một tập gồm p mẫu học, chúng ta lần lượt xây dựng được p hàm giá như vậy. Việc học của mạng hay nhiệm vụ của giải thuật thực chất là tìm kiếm tập trọng số W trong không gian RM (M là số trọng số có trong mạng) để lần lượt tối thiểu hoá các hàm giá như vậy. Điều đáng chú ý là việc tối thiểu hoá được tiến hành liên tiếp nhau và theo chu kỳ đối với các hàm giá.

Để tối thiểu hoá các hàm giá như vậy, giải thuật Lan truyền ngược sai số sử dụng phương pháp giảm gradient để điều chỉnh các trọng số liên kết giữa các nơron. Bản chất của phương pháp này là khi sai số E được vẽ như hàm của tham số gây ra sai số sẽ phải có một cực tiểu tại bộ giá trị nào đó của tham số. Khi quan sát độ dốc của đường cong, chúng ta quyết định phải thay đổi tham số thế nào để có thể tiến gần đến cực tiểu cần tìm kiếm hơn. Trong hình vẽ dưới đây, giá trị của trọng số phải giảm nếu đạo hàm dE/dW là dương

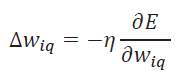
Hình 2.2 Sai số E được xét là hàm của trọng số W

Bằng biểu thức, chúng ta có thể biểu diễn phương pháp giảm gradient như sau :

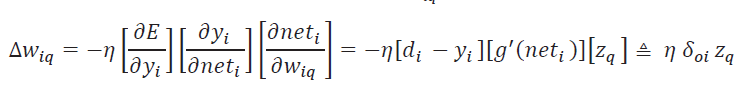
∆w= w(new) – w(old) = -η.∂E/∂w

Ở đây η là hằng số dương xác định tốc độ giảm giá trị của w, còn dấu âm chỉ chiều giảm gradient.

Áp dụng phương pháp giảm gradient đối với các trọng số liên kết giữa các nơron trong lớp ẩn tới các nơron của lớp ra ta có:

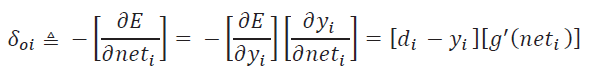


Do hàm sai số E là một hàm phức tạp và là hàm gián tiếp của trọng số wiq, sử dụng nguyên tắc tính đạo hàm gián tiếp cho ta có:



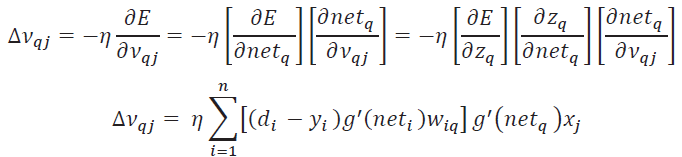
Trong đó là tín hiệu sai số và chỉ số *oi* có nghĩa là nút thứ *i* trong lớp ra.

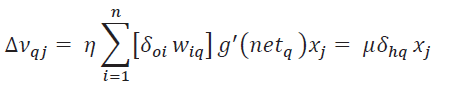
Tín hiệu sai số được tính như sau:



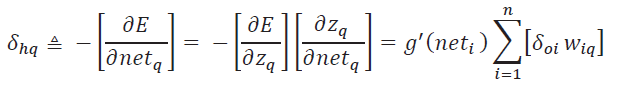
Trong đó neti là tín hiệu của nơron thứ i trên lớp ra và g’(neti) = ∂g(neti)/∂neti.

Để điều chỉnh trọng số của các liên kết giữa lớp vào tới lớp ẩn ta cũng sử dụng phương pháp gradient và lấy đạo hàm theo các biến trung gian như đã áp dụng ở trên. Xét liên kết giữa nơron thứ j ở lớp vào và nơron thứ q ở lớp ra:





Trong đó δhq là tín hiệu lỗi của nơron thứ q trong lớp ẩn và được định nghĩa như sau:



Với netq là tín hiệu vào của nơron thứ q, như vậy tín hiệu lỗi của nơron trên lớp ẩn khác với tín hiệu lỗi của nơron trên lớp ra. Vì sự khác nhau này, thủ tục điều chỉnh trọng số được gọi là luật học delta mở rộng. Nhìn lại công thức tín hiệu lỗi δhq của nơron thứ q trong lớp ẩn được xác định từ các tín hiệu lỗi δoi của các nơron trên lớp ra.

**Tổng quát đối với lớp bất kỳ, luật lan truyền ngược có dạng:**

 trong đó “output\_i” là đầu ra của nơron i và “input\_j” là đầu vào của nơron j, δi là tín hiệu học được định nghĩa trong công thức trên.

**Thuật toán lan truyền ngược sai số được xây dựng như sau:**

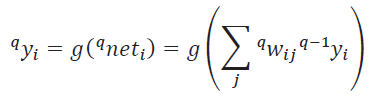
Xét một mạng nơron truyền thẳng có Q lớp, q = 1, 2, …, Q, và gọi neti và yi là tín hiệu vào và ra của nơron thứ i trong lớp q. Mạng này có m đầu vào và n đầu ra. Đặt wij là trọng số của liên kết từ nơron thứ j trong lớp q-1 tới nơron thứ i trong lớp q.

**Đầu vào:** Một tập các cặp mẫu học {(x(k), d(k)) | k= 1, 2, …, p}

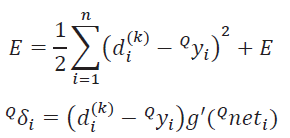
* Bước 1 (khởi tạo): Chọn một hằng số η > 0 và Emax (dung sai cho phép). Khởi tạo ngẫu nhiên các trọng số wij trong khoảng giá trị nhỏ. Đặt E =0 và k = 1.
* Bước 2 (thực hiện một quá trình lặp cho việc huấn luyện mạng)
* Sử dụng mẫu học thứ k ;
* Tại lớp vào (q =1), với mọi i ta có: qyi = 1yi = x(k)

i

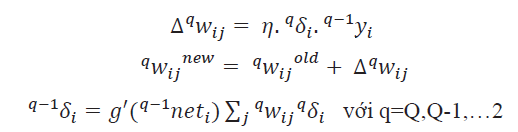
* Bước 3 (lan truyền tín hiệu từ lớp vào tới lớp ra)



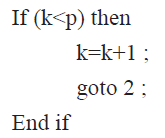
* Bước 4 (xác định tín hiệu lỗi Qδi tại lớp ra)



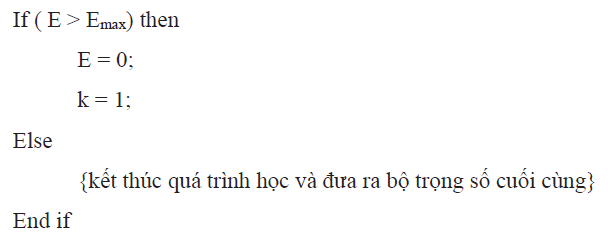
* Bước 5 (lan truyền ngược sai số): Lan truyền ngược sai số để điều chỉnh các trọng số và tính toán tín hiệu lỗi q-1δi cho các lớp trước:



* Bước 6 (kiểm tra điều kiện lặp)

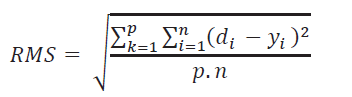


* Bước 7 (kiểm tra lỗi tổng cộng hiện thời đã chập nhận được chưa)



Mỗi lần toàn bộ tập mẫu học được lan truyền qua mạng được gọi là một epoch. Số epoch phụ thuộc vào từng trường hợp cụ thể và sự khởi tạo ban đầu. Có trường hợp thuật toán phải sau hàng chục nghìn epoch mới hội tụ tới lời giải. Nếu tham số khởi tạo không phù hợp có thể làm cho quá trình học không đạt kết quả mong muốn.

Đối với mỗi epoch ta tính sai số trung bình của mạng theo công thức sau:



Trong đó p là số mẫu được dùng để luyện mạng, n là số biến của véc-tơ đầu ra. Sai số RMS thường được dùng để đánh giá kết quả học của mạng nơron.

### 2.1.2. Một số yếu tố ảnh hưởng đến quá trình học theo phương pháp lan truyền ngược

**Khởi tạo các trọng số**

Các giá trị được khởi tạo ban đầu cho các trọng số trong mạng lan truyền ngược sai số ảnh hưởng rất lớn đến kết quả học cuối cùng của mạng. Các giá trị này thường được khởi tạo ngẫu nhiên trong phạm vi giá trị tương đối nhỏ. Thông thường hàm truyền sử dụng cho mạng MLP là hàm sigmoid, do vậy nếu ta chọn các giá trị trọng số khởi tạo lớn thì các hàm này có thể bão hoà ngay từ đầu và dẫn tới hệ thống có thể bị tắc ngay tại một cực tiểu địa phương hoặc tại một vùng bằng phẳng nào đó gần điểm xuất phát.

**Hằng số học**

Hằng số học η cũng là một yếu tố quan trọng ảnh hưởng đến hiệu quả và độ hội tụ của giải thuật lan truyền ngược sai số. Không có hằng số η phù hợp cho tất cả các bài toán khác nhau. Hằng số học này thường được chọn bằng thực nghiệm cho mỗi bài toán ứng dụng cụ thể bằng phương pháp thử sai.

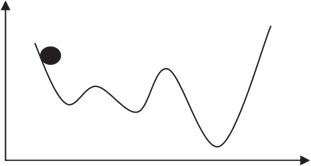
Trong nhiều ứng dụng thực tế cho thấy một hằng số học có thể phù hợp ở thời điểm bắt đầu của quá trình học nhưng lại không phù hợp với giai đoạn sau của quá trình học. Do đó, có một phương pháp hiệu quả hơn đó là sử dụng hằng số học thích nghi. Một cách xử lý đơn giản cho vấn đề này đó là kiểm tra xem các trọng số mới có làm giảm hàm giá hay không, nếu không thì có thể các trọng số đã vượt quá xa vùng cực tiểu và như vậy hằng số η cần phải giảm. Trái lại, nếu sau vài vòng lặp, hàm giá liên tục giảm thì ta có thể thử tăng hằng số η để đẩy nhanh hơn tốc độ hội tụ đến giá trị cực tiểu.

**Hằng số quán tính**

Tốc độ học của giải thuật Lan truyền ngược sai số có thể rất chậm nếu hằng số học nhỏ, nhưng nếu hằng số học lớn thì nó lại có thể gây ra sự dao động lớn trong quá trình tìm giá trị cực tiểu theo phương pháp giảm gradient. Để giải quyết vấn đề này người ta thường thêm thành phần quán tính vào các phương trình hiệu chỉnh trọng số như sau:

∆w(t) = -ηVE(t) + α ∆w(t-1) với α là hằng số quán tính, α E [0, 1]

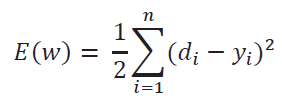
Nhờ thành phần này, quá trình học có thể vượt qua điểm cực tiểu địa phương để tìm đến điểm cực tiểu toàn cục, đồng thời thành phần quán tính cũng ngăn cản sự thay đổi đột ngột của các trọng số theo hướng khác với hướng mà lời giải đang di chuyển đến.



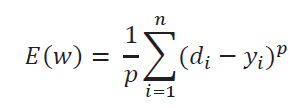
Hình 2.3 Minh họa về ý nghĩa của quán tính trong thực tế

**Hàm giá**

Trong phần nghiên cứu trên hàm giá được chọn là hàm bình phương sai số:



Tuy nhiên, nó có thể được thay thế bằng một hàm F(y,d) bất kỳ có đạo hàm và đạt cực tiểu khi hai đối số di và yi bằng nhau. Thông thường hàm giá được chọn có dạng:



Với 1 ≤ p ≤∞, như vậy khi chọn p = 2 ta có hàm giá là hàm bình phương sai số như đã xét ở trên.

## 2.2. Giải thuật di truyền [2]

Từ trước tới nay, trong các nghiên cứu và ứng dụng tin học đã xuất hiện nhiều bài toán chưa tìm ra được phương pháp giải nhanh và hợp lý. Phần lớn đó là các bài toán tối ưu nảy sinh trong các ứng dụng. Để giải các bài toán này người ta thường phải tìm đến một giải thuật hiệu quả mà kết quả thu được chỉ là xấp xỉ tối ưu. Trong nhiều trường hợp chúng ta có thể sử dụng giải thuật xác suất, tuy không bảo đảm kết quả tối ưu nhưng cũng có thể chọn các giá trị sao cho sai số đạt được sẽ nhỏ như mong muốn.

Theo lời giải xác suất, việc giải bài toán quy về quá trình tìm kiếm trên không gian tập hợp các lời giải có thể. Tìm được lời giải tốt nhất và quá trình được hiểu là tối ưu. Với miền tìm kiếm nhỏ, một số thuật toán cổ điển được sử dụng. Tuy nhiên đối với các miền lớn, phải sử dụng các kỹ thuật trí tuệ nhân tạo đặc biệt, giải thuật Di truyền là một trong những công cụ đó. Ý tưởng của giải thuật Di truyền là mô phỏng những gì mà tự nhiên đã thực hiện. Giải thuật Di truyền hình thành dựa trên quan niệm cho rằng: quá trình tiến hóa tự nhiên là quá trình hoàn hảo nhất, hợp lý nhất và tự nó đã mang tính tối ưu. Quá trình tiến hoá thể hiện tính tối ưu ở chỗ thế hệ sau bao giờ cũng tốt hơn thế hệ trước.

Giải thuật Di truyền áp dụng quá trình tiến hóa tự nhiên để giải các bài toán tối ưu trong thực tế (từ tập các lời giải có thể ban đầu thông qua nhiều bước tiến hóa hình thành các tập hợp mới với lời giải tốt hơn và cuối cùng sẽ tìm được lời giải gần tối ưu).

Giải thuật Di truyền là một kỹ thuật của khoa học máy tính nhằm tìm kiếm giải pháp thích hợp cho các bài toán tối ưu tổ hợp (combinatorial optimization). Giải thuật Di truyền là một phân ngành của giải thuật tiến hóa vận dụng các nguyên lý của tiến hóa như di truyền, đột biến, chọn lọc tự nhiên, và trao đổi chéo.

Giải thuật Di truyền thường được ứng dụng nhằm sử dụng ngôn ngữ máy tính để mô phỏng quá trình tiến hoá của một tập hợp những đại diện trừu tượng (gọi là những nhiễm sắc thể) của các giải pháp có thể (gọi là những cá thể) cho bài toán tối ưu hóa vấn đề. Tập hợp này sẽ tiến triển theo hướng chọn lọc những giải pháp tốt hơn.

Thông thường, những giải pháp được thể hiện dưới dạng nhị phân với những chuỗi 0 và 1, nhưng lại mang nhiều thông tin mã hóa khác nhau. Quá trình tiến hóa xảy ra từ một tập hợp những cá thể hoàn toàn ngẫu nhiên ở tất cả các thế hệ. Trong từng thế hệ, tính thích nghi của tập hợp này được ước lượng, nhiều cá thể được chọn lọc định hướng từ tập hợp hiện thời (dựa vào thể trạng), được sửa đổi (bằng đột biến hoặc tổ hợp lại) để hình thành một tập hợp mới. Tập hợp này sẽ tiếp tục được chọn lọc lặp đi lặp lại trong các thế hệ kế tiếp của giải thuật.

### 2.2.1. Các khái niệm cơ bản

Giải thuật di truyền dựa vào quá trình tiến hoá trong tự nhiên nên các khái niệm và thuật ngữ của nó đều có liên quan tới các thuật ngữ của di truyền học.

#### 2.2.1.1. Cá thể, nhiễm sắc thể

Một cá thể trong giải thuật Di truyền, biểu diễn một giải pháp của bài toán. Tuy nhiên không giống với trong tự nhiên, một cá thể có nhiều nhiễm sắc thể (NST), có 1 thì gọi là thể đơn bội, còn nếu có nhiều thì là thể đa bội, ở đây để giới hạn trong giải thuật Di truyền ta quan niệm một cá thể có một nhiễm sắc thể. Do đó khái niệm cá thể và nhiễm sắc thể trong giải thuật Di truyền coi như là tương đương.

Một NST được tạo thành từ nhiều gen, mỗi gen có thể có các giá trị khác nhau để quy định một tính trạng nào đó. Trong giải thuật Di truyền, một gen được coi như một phần tử trong chuỗi NST.

#### 2.2.1.2. Quần thể

Quần thể là một tập hợp các cá thể có cùng một số đặc điểm nào đấy. Trong giải thuật Di truyền ta quan niệm quần thể là một tập các lời giải của một bài toán.

#### 2.2.1.3. Các toán tử di truyền

* **Chọn lọc**: Trong tự nhiên, quá trình chọn lọc và đấu tranh sinh tồn đã làm thay đổi các cá thể trong quần thể. Những cá thể tốt, thích nghi được với điều kiện sống thì có khả năng đấu tranh lớn hơn, do đó có thể tồn tại và sinh sản. Các cá thể không thích nghi được với điều kiện sống thì dần mất đi. Dựa vào nguyên lý của quá trình chọn lọc và đấu tranh sinh tồn trong tự nhiên, chọn lựa các cá thể trong giải thuật Di truyền chính là cách chọn các cá thể có độ thích nghi tốt để đưa vào thế hệ tiếp theo hoặc để cho lai ghép, với mục đích là sinh ra các cá thể mới tốt hơn. Có nhiều cách để lựa chọn nhưng cuối cùng đều nhằm đáp ứng mục tiêu là các cá thể tốt sẽ có khả năng được chọn cao hơn.
* **Lai ghép**: Lai ghép trong tự nhiên là sự kết hợp các tính trạng của bố mẹ để sinh ra thế hệ con. Trong giải thuật di truyền, lai ghép được coi là một sự tổ hợp lại các tính chất (thành phần) trong hai lời giải cha mẹ nào đó để sinh ra một lời giải mới mà có đặc tính mong muốn là tốt hơn thế hệ cha mẹ. Đây là một quá trình xảy ra chủ yếu trong giải thuật Di truyền.
* **Đột biến**: Đột biến là một sự biến đổi tại một (hay một số) gen của nhiễm sắc thể ban đầu để tạo ra một nhiễm sắc thể mới. Đột biến có xác suất xảy ra thấp hơn lai ghép. Đột biến có thể tạo ra một cá thể mới tốt hơn hoặc xấu hơn cá thể ban đầu. Tuy nhiên trong giải thuật Di truyền thì ta luôn muốn tạo ra những phép đột biến cho phép cải thiện lời giải qua từng thế hệ.

### 2.2.2. Mô hình giải thuật di truyền

Với các khái niệm được nêu ở trên, giải thuật Di truyền được mô tả như sau:

1. **[Bắt đầu]** Nhận các tham số cho thuật toán.
2. **[Khởi tạo]** Sinh ngẫu nhiên một quần thể gồm n cá thể (là n lời giải cho bài toán).
3. [**Quần thể mới**] Tạo quần thể mới bằng cách lặp lại các bước sau cho đến khi quần thể mới hoàn thành.
   1. **[Thích nghi**] Ước lượng độ thích nghi *eval(x)* của mỗi cá thể.
   2. [**Kiểm tra]** Kiểm tra điều kiện kết thúc giải thuật.
   3. [**Chọn lọc**] Chọn hai cá thể bố mẹ từ quần thể cũ theo độ thích nghi của chúng (cá thể có độ thích nghi càng cao thì càng có nhiều khả năng được chọn).
   4. [**Lai ghép**] Với một xác suất lai ghép được chọn, lai ghép hai cá thể
   5. bố mẹ để tạo ra một cá thể mới.
   6. [**Đột biến**] Với một xác suất đột biến được chọn, biến đổi cá thể mới.
4. [**Chọn kết quả**] Nếu điều kiện dừng được thỏa mãn thì thuật toán kết thúc và trả về lời giải tốt nhất trong quần thể hiện tại.

### 2.2.3. Áp dụng giải thuật Di truyền cho bài toán mạng nơron nhân tạo

Để có thể sử dụng được giải thuật Di truyền vào việc học của mạng nơron cần phải thực hiện một số bước như sau:

* Xây dựng hàm giá: Hàm giá này sẽ được sử dụng để tạo nên độ phù hợp của các cá thể và của cả quần thể trong GA. Trong nghiên cứu này sử dụng hàm sai số RMS tương tự như trong giải thuật Lan truyền ngược sai số.
* Mã hoá nhiễm sắc thể: Mỗi cá thể trong GA sẽ thay mặt cho một bộ trọng số của mạng nơron. Ở đây ta không cần phải phân biệt trọng số nào ở lớp nào mà ta chỉ cần trải tất cả các trọng số lên sơ đồ gen của nhiễm sắc thể.
* Thực hiện giải thuật Di truyền:
* Chọn lọc: Gán 1 nhiễm sắc thể con bằng một nhiễm sắc có hàm giá nhỏ hơn của 2 nhiễm sắc thể bố mẹ ngẫu nhiên (hoặc là bố, hoặc là mẹ).
* Lai ghép: Toán tử lai ghép này sẽ đưa một giá trị vào mỗi vị trí của nhiễm sắc thể con bằng cách lấy ngẫu nhiên một giá trị tại cùng vị trí của nhiễm sắc thể cha hoặc mẹ.
* Đột biến: Một gen (trọng số) được lựa chọn ngẫu nhiên với một xác suất pmutation để tiến hành đột biến sử dụng phương pháp BIASED: với mỗi genđược chọn đột biến nó sẽ được cộng thêm một giá trị ngẫu nhiên quanh giá trị gốc ban đầu.

## 2.3. Kết hợp các giải thuật

Như chúng ta đã biết sử dụng giải thuật Lan truyền ngược sai số để tối ưu hoá trọng số của mạng nơron nhân tạo đang được sử dụng rộng rãi hiện nay. Tuy nhiên, giải thuật này hoạt động theo cơ chế giảm gradient nên nó khó có thể tìm ra được cực trị toàn cục. Trong nghiên cứu này sử dụng giải thuật Di truyền để tối ưu hoá trọng số của mạng giúp quá trình học của mạng được tốt hơn.

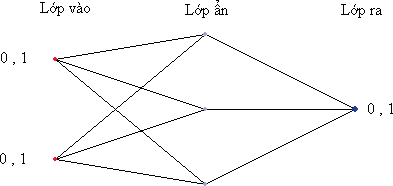
Thử nghiệm hai phương pháp với bài toán nổi tiếng XOR. Có 4 mẫu học như sau:

Bảng 2.1 Giá trị đầu vào và ra của bài toán XOR

|  |  |  |
| --- | --- | --- |
| X1 | X2 | Y |
| 0 | 0 | 0 |
| 0 | 1 | 1 |
| 1 | 0 | 1 |
| 1 | 1 | 0 |

Tham số chung cho cả hai phương pháp:

* Mạng nơron sử dụng là mạng có một lớp ẩn
* Số nơron trong lớp ẩn: 3
* Ngưỡng dừng lặp: 1000 vòng lặp

*Hình 2.4 Mô hình mạng nơron cho bài toán XOR*

Hình 2.4: Tham số của giải thuật lan truyền ngược sai số:

* Hằng số học: 0.3

Tham số của giải thuật di truyền:

* Số lượng quần thể: 100
* Xác suất lai: 0.3
* Xác suất đột biến: 0.1
* Kết quả trong 1.000.000 lần chạy với sai số dừng 0.05:
* Giải thuật Di truyền chạy thành công được 837.532 lần, trung bình 19 vòng lặp một lần chạy.
* Giải thuật Lan truyền ngược sai số chạy thành công được 763.908 lần trung bình 214 vòng lặp.

Ta có thể thấy rằng giải thuật Di truyền có khả năng đạt được yêu cầu về hội tụ (sai số ≤ 0.05) tức tìm vùng chứa cực trị toàn cục dễ dàng hơn so với giải thuật Lan truyền ngược sai số. Hay nói cách khác giải thuật Lan truyền ngược sai số dễ rơi vào vùng chứa cực tiểu cục bộ hơn giải thuật di truyền.

* Kết quả trong 1.000.000 lần chạy với sai số dừng 0.02:
* Giải thuật Di truyền chạy thành công được 402.466 lần trung bình 51 vòng lặp một lần chạy.
* Giải thuật Lan truyền ngược sai số chạy thành công được 567.466 lần trung bình 202 vòng lặp.

Mặc dù giải thuật Di truyền có khả năng đạt tới cực trị toàn cục cho quá trình tìm kiếm nhưng do có kết hợp những yếu tố ngẫu nhiên nên tốc độ tìm kiếm nói chung là rất chậm. Mặt khác nó không thể hoàn toàn đạt được tới cực trị toàn cục mà chỉ cho những kết quả xung quanh đó. Đối lập với GA, giải thuật Lan truyền ngược sai số lại cho phép đạt được những cực trị nếu như điểm xuất phát của quá trình tìm kiếm nằm trong vùng cực trị toàn cục.

Từ 2 kết quả trên ta có nhận xét: giải thuật Di truyền có thể đạt đến vùng chứa cực tiểu toàn cục (sai số 0.05) dễ dàng hơn so với Giải thuật Lan truyền ngược sai số. Tuy nhiên, để đạt đến chính xác vị trí cực tiểu toàn cục (sai số nhỏ dần) thì giải thuật Di truyền lại rất kém. Trong khi đó, hầu hết các trường hợp Giải thuật Lan truyền ngược sai số khi đã đưa mạng đến được vùng chứa cực tiểu toàn cục (sai số 0.05) thì Giải thuật Lan truyền ngược sai số sẽ đưa mạng đến chính xác cực tiểu toàn cục. Do đó, việc kết hợp giải thuật Di truyền và Giải thuật Lan truyền ngược sai số có nhiều cơ hội đưa mạng đến được chính xác cực tiểu toàn cục.

Ta thấy rằng có thể kết hợp giải thuật Di truyền và Giải thuật Lan truyền ngược sai số nhằm nâng cao hiệu quả của BP. Giải thuật Di truyền sẽ khoanh vùng chứa cực tiểu toàn cục của hàm lỗi, sau đó Giải thuật Lan truyền ngược sai số xuất phát từ bộ trọng số này để tiến đến cực tiểu toàn cục.

Có nhiều cách để kết hợp giải thuật Di truyền vào mạng nơron nhưng cách đơn giản và khá hiệu quả là ta thực hiện lai ghép hai giải thuật nối tiếp nhau.

Tập trọng số được mã hoá thành các nhiễm sắc thể và được tiến hoá nhờ GA. Kết thúc quá trình tiến hoá, bộ trọng số tốt nhất tương ứng với cá thể ưu việt nhất trong quần thể được lựa chọn làm những trọng số khởi tạo cho giải thuật BP. Nó chính là bộ tham số cho phép xác định điểm gần cực trị nhất của hàm giá.

Với việc lai ghép này, giải thuật Lan truyền ngược sai số lược bỏ đi một số bước sau:

* Không khởi tạo các giá trị trọng số ban đầu vì tập trọng số đã được lấy từ kết quả của giải thuật di truyền.
* Thành phần quán tính trong các phương trình hiệu chỉnh trọng số là không cần thiết vì tập trọng số xuất phát đã khá gần lời giải; tác dụng chống dao động và thay đổi đột ngột các trọng số theo hướng khác với hướng của lời giải trở nên không cần thiết.

Việc thử nghiệm giải thuật kết hợp này được tiến hành với bài toán XOR ngưỡng sai số mong muốn là 0.001. Các tham số cũng như các phần trước. Giải thuật gồm hai bước chính:

* GA sẽ đưa mạng đạt đến sai số 0.05;
* BP sẽ nhận bộ trọng số tốt nhất của giải thuật Di truyền đóng vai trò là trọng số khởi tạo (có sai số 0.05) để đưa mạng đến sai số mong muốn 0.001.

Ta có thể tóm tắt khả năng hội tụ của mạng với hai phương pháp học: giải thuật Di truyền và giải thuật Lan truyền ngược sai số khi ngưỡng sai số dừng lặp khác nhau trong bảng 2.2.

Bảng 2.2 So sánh khả năng hội tụ của mạng khi sử dụng hai phương pháp học GA và BP với sai số dừng lặp khác nhau

|  |  |  |  |
| --- | --- | --- | --- |
| Sai số dừng lặp | Số lần hội tụ trong 1.000.000 lần thử nghiệm | | |
| GA | BP | Kết hợp |
| 0.05 | 83,75% | 75,64% |  |
| 0.02 | 40,25% | 56,75% |  |
| 0.001 | 0% | 18,44% | 60,45% |

So sánh với việc sử dụng GA và BP riêng rẽ thì giải thuật kết hợp này cho kết quả tốt hơn rất nhiều.

# TÀI LIỆU THAM KHẢO

1. Blum, E. K. and L. K. Li. Approximation Theory and feedforward networks, Neural Networks, 1991, Vol. 4, pp. 511-515.
2. D.E. Goldberg, Genetic Algorithm in Search, Optimization and Machine Learning, Addison Wesley, Reading, MA, (1989).
3. Hecht-Nielsen, R. 1989. Theory of backpropagation neural network, In Proceedings of the International Ioint Conference on Neural Networks, Washington DC., (June 1989), IEEE TAB Neural Network Committee, pp. I593-P605.
4. Kolmogorov A. N. On the representation of continuous functions of many variables by superposition of continuous function of one variable and addition, Dokl, Akad, Nauk SSSR, 114, 953-956, Trans. Am. Math-Soc. 2(28), 55-59, (1957).
5. Lê Minh Trung. Giáo trình mạng neuron nhân tạo, Nhà xuất bản thống kê, (1999).
6. Martin T. Hagan, Neural Network Design, PWS Publishing Company, (1996).
7. Oscar R. Dolling, Eduardo A. Varas, Artificial neural networks for stream flow prediction, Journal of Hydraaulic research, 40(5), 547-554, (2002).
8. Phạm Thị Hoàng Nhung, Nghiên cứu ứng dụng các phương pháp học máy tiên tiến trong công tác dự báo, vận hành hồ Hòa Bình, Luận văn Thạc sỹ, (2007).