ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

ĐẶNG QUỐC HÙNG

DỰ ĐOÁN SỰ TƯƠNG TÁC GIỮA CÁC PROTEIN DỰA TRÊN KỸ THUẬT HỌC SÂU

LUẬN VĂN THẠC SĨ
Ngành Công nghệ thông tin

HÀ NỘI - 2017

ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

ĐẶNG QUỐC HÙNG

DỰ ĐOÁN SỰ TƯƠNG TÁC GIỮA CÁC PROTEIN DỰA TRÊN KỸ THUẬT HỌC SÂU

Ngành: Công nghệ thông tin
Chuyên ngành: Kỹ thuật phần mềm
Mã số: 60480103

TÓM TẮT LUẬN VĂN THẠC SĨ
Ngành Công nghệ thông tin

HÀ NỘI - 2017

LỜI CẢM ƠN

Đầu tiên, em xin gửi lời cảm ơn chân thành và sâu sắc nhất đến thầy Đặng Thanh Hải, người đã trực tiếp hướng dẫn, chỉ bảo tận tình, giúp đỡ em trong suốt quá trình học tập, nghiên cứu và thực hiện đề tài này.

Em cũng xin bày tỏ lòng biết ơn sâu sắc đến các Thầy Cô giảng viên và cán bộ trong Khoa Công nghệ thông tin nói riêng và trong trường Đại học Công nghệ - Đại học Quốc Gia Hà Nội nói chung, đã dành hết tâm huyết, tận tình hướng dẫn học viên chúng em trong suốt quãng thời gian qua.

Em xin cảm ơn Khoa Công nghệ thông tin đã tạo điều kiện cho chúng em học tập trong môi trường nghiên cứu lành mạnh và thuận lợi để chúng em có thể phát triển được niềm đam mê của mình.

Mình cũng xin gửi lời cảm ơn tới bạn Trác Quang Thịnh vì sự hỗ trợ của bạn trong suốt thời gian nghiên cứu.

Cuối cùng, tôi xin gửi lời cảm ơn tới các bạn trong trường đã ủng hộ và giúp đỡ tôi trong suốt quá trình học tập và thực hiện đề tài.

Hà Nội, ngày 12 tháng 10 năm 2017

Học viên

Đặng Quốc Hùng

LỜI CAM ĐOAN

Em xin cam đoan các phương pháp và kỹ thuật sử dụng trong nghiên cứu sự tương tác giữa các protein dựa trên kĩ thuật học sâu được trình bày trong luận văn này là do em thực hiện dưới sự hướng dẫn của Thầy Đặng Thanh Hải. Tất cả những tham khảo từ các nghiên cứu liên quan đều được trích dẫn nguồn gốc rõ ràng từ danh mục tài liệu tham khảo trong luận văn.

Trong luận văn này, không có việc sao chép tài liệu, các công trình nghiên cứu của người khác mà không ghi rõ trong tài liệu tham khảo. Nếu phát hiện có bất kì sự gian lận nào, em xin hoàn toàn chịu trách nhiệm trước hội đồng cũng như kết quả luận văn của mình.

Hà Nội, ngày 12 tháng 10 năm 2017

Học viên

Đặng Quốc Hùng

MỤC LỤC

DANH MỤC CÁC KÝ HIỆU VÀ CHỮ VIẾT TẮT

STT	**Tiếng Anh**	**Tiếng Việt**
1	Activation function	Hàm kích hoạt
2	Area under the curve (AUC)	Diện tích dưới đường cong
3	Convolutional layer	Lớp tích chập
4	Convolutional Neural Networks (CNNs)	Mạng nơ ron tích chập
5	Distribution	Phân phối
6	Feature map	Lớp ánh xạ đặc trưng
7	Filter	Bộ lọc
8	Fully connected	Kết nối đầy đủ
9	Kernel	Hàm nhân
10	K-fold cross validation	Kiểm định chéo k-fold
11	Layer	Lớp/tầng
12	Linear	Tuyến tính
13	Overfitting	Quá vừa dữ liệu
14	Quasi Sequence Order (QSO)	Trình tự Quasi
15	Stride	Bước trượt
16	Support vector machine (SVM)	Máy véc tơ hỗ trợ
17	Threshold	Ngưỡng
18	Protein - Protein interactions(PPIs)	Sự tương tác protein

DANH MỤC CÁC HÌNH VẼ

DANH MỤC CÁC BẢNG

MỞ ĐẦU

Protein hay còn gọi là chất đạm là những đại phân tử được cấu tạo theo nguyên tắc đa phân mà các đơn phân là amino axit. Amino axit được cấu tạo bởi ba thành phần: một là nhóm amin (-NH2), hai là nhóm cacboxyl (-COOH) và cuối cùng là nguyên tử cacbon trung tâm đính với một nguyên tử hyđro và nhóm biến đổi R quyết định tính chất của amino axit [55]. Protein và amino axit đóng vai trò vô cùng quan trọng trong các cơ thể sống, bao gồm: là nền tảng tạo nên sức sống của cơ thể, tham gia quá trình trao đổi chất dinh dưỡng, điều hoà nước, là chất bảo vệ cơ thể, cung cấp năng lượng cho các hoạt động sống.

Tương tác protein là quá trình tác động qua lại giữa các protein với nhau hoặc giữa các protein với các phân tử khác trong tế bào. Sự tương tác này tác động tới các hoạt động của tế bào và ảnh hưởng đến quá trình sống của các cơ thể sống. Protein và tương tác giữa các protein là trung tâm của hầu hết các quá trình sinh học. Thông thường, protein hiếm khi hoạt động một cách độc lập mà thực hiện chức năng của chúng thông qua sự tương tác với các đơn vị phân tử sinh học khác. Do đó, việc kiểm tra các tương tác protein-protein (PPI) là cần thiết để hiểu các cơ chế phân tử của các quá trình sinh học, dự đoán tương tác giữa các protein là bài toán quan trọng và là ưu tiên của ngành khoa học sinh học. Tương tác protein thường được xác định bằng các phương pháp lý hóa sinh, là các phương pháp nghiên cứu thực nghiệm trong các phòng thí nghiệm. Tuy nhiên, các phương pháp thực nghiệm này rất đắt tiền, mất thời gian và có tỉ lệ thành công thấp. Do đó, việc phát triển các mô hình tính toán đáng tin cậy tạo thuận lợi cho việc xác định các tương tác protein có ý nghĩa thực tiễn to lớn giúp phát hiện các tương tác protein có khả năng xảy ra cao làm tiền đề cung cấp tập lựa chọn ưu tiên cho thực nghiệm nhằm đem lại khả năng thành công cao hơn và tiết kiệm chi phí thực nghiệm.

Những lý do này đã thúc đẩy sự phát triển mạnh của hướng nghiên cứu tin sinh học. Một số lượng lớn các phương pháp tính toán đã được đề xuất để dự đoán các tương tác Proteins.

Nội dung luận văn được chia thành 3 chương như sau:

Chương 1 giới thiệu một cách sơ lược về bài toán dự đoán sự tương tác giữa các proteins cũng như các mô hình dự đoán tương ứng điển hình, gần đây nhất.

Chương 2 trình bày các kiến thức cơ bản về mạng nơ ron, kỹ thuật học sâu và đi sâu vào mạng nơ ron tích chập.

Chương 3 trình bày về xây dựng mô hình để dự đoán, các kết quả đạt được của mô hình đồng thời đi so sánh các kết quả của các phương pháp khác, qua đó có cái nhìn tổng quan về chất lượng của các phương pháp này.

CHƯƠNG 1. TỔNG QUAN VỀ DỰ ĐOÁN TƯƠNG TÁC PROTEINS

1.1. Giới thiệu về tương tác giữa các proteins

Protein là những đại phân tử được cấu tạo theo nguyên tắc đa phân mà các đơn phân là axít amin. Chúng kết hợp với nhau thành một mạch dài nhờ các liên kết peptide (gọi là chuỗi polypeptide). Các chuỗi này có thể xoắn cuộn hoặc gấp theo nhiều cách để tạo thành các bậc cấu trúc không gian khác nhau của protein. Trong tế bào động vật, protein có vai trò hết sức quan trọng. Chúng tham gia cấu trúc tế bào, là những enzym xúc tác cho các quá trình sinh lý sinh hóa xảy ra trong tế bào. Protein còn tham gia vào các quá trình vận chuyển, bảo vệ, điều khiển, là nơi dự trữ chất dinh dưỡng, nhận biết các loại phân tử khác nhau, chịu trách nhiệm về sự vận động của cơ thể sống ở mức tế bào và cơ thể. Các chức năng này có thể do một hoặc nhiều phân tử protein đặc hiệu đảm nhiệm.

Tương tác protein là quá trình tác động qua lại giữa các protein với nhau hoặc giữa các protein với các phân tử khác trong tế bào. Sự tương tác này tác động tới các hoạt động của tế bào và ảnh hưởng đến quá trình sống của các cơ thể sống.

Dựa vào đặc điểm chức năng và cấu trúc PPIs có thể được phân loại theo nhiều cách khác nhau theo bề mặt tương tác thì PPIs có thể là Homo hoặc là hetero-oligomeric, theo sự ổn định có thể phân loại thành: bắt buộc hoặc không bắt buộc, theo mức độ bền vững thì có thể phân loại thành: PPIs yếu và PPIs bền. Một tương tác cũng có thể được phân loại khác nhau trong các điều kiện khác nhau. Ví dụ, PPI có thể là tương tác yếu trong cơ thể (in Vivo) nhưng cũng có thể trở thành tương tác bền trong các điều kiện nhất định trong tế bào. Theo quan sát, các protein hiếm khi thực hiện chức năng của chúng một mình mà thường kết hợp với các protein khác bằng cách hình thành một mạng tương tác protein protein khổng lồ. Tập hợp các tương tác của protein gọi là mạng tương tác protein – protein (protein – protein interactions - PPIs). Việc tạo bản đồ tương tác PPIs không chỉ cung cấp cái nhìn sâu sắc hơn về chức năng của protein mà còn giúp làm rõ các cơ chế phân tử trong tế bào. Nghiên cứu PPIs là bước cơ bản để tìm hiểu chức năng của protein trong tế bào. Theo Phizicky và Fields, PPIs có thể làm thay đổi tính chất của các enzymes, tạo ra một vị trí liên kết mới, ngừng hoạt động hoặc phá hủy một protein hoặc có thể dẫn đến thay đổi đặc tính của protein.

Mạng PPIs có thể được định nghĩa là một hệ thống phức tạp các proteins được liên kết bởi các tương tác giữa chúng. Mạng PPIs thường được biểu diễn bằng đồ thị bao gồm các đỉnh và các cạnh, trong đó các proteins biểu diễn các đỉnh của đồ thị, các cạnh biểu diễn bởi tương tác giữa các proteins. Căn cứ vào đồ thị này, các phương pháp tính toán khác nhau như khai phá dữ liệu, học máy, phương pháp thống kê có thể được thiết kế để tổ chức các mạng PPI ở các cấp độ khác nhau. Việc kiểm tra các mô hình đồ thị của mạng có thể mang lại nhiều hiểu biết, ví dụ các proteins láng giềng

trong đồ thị có thể chia sẻ các chức năng với nhau. Vì vậy, các chức năng của protein có thể được dự đoán bởi việc quan sát các proteins khi chúng tương tác với nhau và protein đó thuộc về phức hợp proteins nào. Ngoài ra, các đồ thị con được kết nối trong mạng có thể xem như là các phức hợp proteins có chức năng như một đơn vị trong quá trình sinh học cụ thể. Việc tìm hiểu các đặc trưng hình học của mạng cũng có thể giúp nâng cao hiểu biết về hệ thống sinh học.

Việc nghiên cứu các mạng PPIs giúp hiểu nhiều hơn về các chức năng, các quá trình, và tổ chức trong tế bào như:

- Dự đoán chức năng Protein: ứng dụng cơ bản nhất của các mạng PPI là phân tích cấu trúc hình học của mạng để dự đoán chức năng Protein. Các modules chức năng được tạo ra có tác dụng như một bộ khung để dự đoán các chức năng của các protein chưa biết. Mỗi module được tạo ra có thể chứa một vài protein chưa được xác định các đặc trưng. Bằng cách kết hợp các protein chưa biết với các protetin đã biết có thể thấy rằng các protein này tham gia một cách tích cực trong việc thực hiện các chức năng đã được chỉ định tới các modules.
- Phân tích gây chết: phân tích cấu trúc hình học của mạng PPI có thể giúp đánh giá mức độ quan trọng của các cạnh và các nút trong mạng. Gây chết là một nhân tố quan trọng trong việc mô tả các đặc tính sinh học của một loại Protein, được xác định bằng cách kiểm tra xem một module chức năng có bị phá vỡ khi protein bị loại bỏ hay không. Các thông tin về gây chết được đưa vào hầu hết các cơ sở dữ liệu PPI.
- Đánh giá khả năng nghiên cứu thuốc từ cấu trúc hình học của mạng. Dự án gen người và các nỗ lực trong việc tìm ra các phương pháp điều trị các căn bệnh của con người là vấn đề quan trọng và cấp thiết của khoa học. Sự công hiệu, đặc tính và tác dụng phụ của các loại thuốc được điều chế tốt hay không phần lớn phụ thuộc vào việc chọn mục tiêu dược phù hợp (pharmacological target). Vì vậy việc xác định các mục tiêu phân tử là bước đầu tiên và quan trọng trong quá trình điều chế thuốc. Mục tiêu của quá trình này là để thu được tập phân tử sinh học đủ nhỏ phục vụ cho việc nghiên cứu, phát triển và thử nghiệm lâm sàng. Các mục tiêu dược có thể là DNA, Lipid, hoặc các chất chuyển hóa. Tuy nhiên trong thực tế thì mục tiêu dược chủ yếu là các proteins.

Tuy nhiên, việc phân tích các mạng PPIs gặp khó khăn như:

- Các tương tác protein thì không đáng tin cậy. Các thí nghiệm nhận được số lượng lớn các dương tính giả. Ví dụ, trong các thí nghiệm dùng phương pháp men hai-lai (Y2H) thì chỉ đạt độ tin cậy xấp xỉ 50%, ngoài ra cũng có nhiều giá trị âm tính giả trong các mạng PPI hiện đang được nghiên cứu.

- Một protein có thể có nhiều chức năng khác nhau. Một protein có thể ở trong một hoặc nhiều nhóm chức năng. Do đó, các cụm chéo nhau nên được xác định trong

các mạng PPI, trong khi các phương pháp phân cụm thông thường tạo ra các cụm tách rời nhau từng đôi một và không có hiệu quả khi áp dụng cho các mạng PPI.

- Hai proteins có các chức năng khác nhau thường tương tác với nhau. Sự liên kết ngẫu nhiên giữa các proteins trong các nhóm chức năng khác nhau làm tăng sự phức tạp về cấu trúc hình học của các mạng PPI

Nghiên cứu PPI đặt ra nhiều thách thức do sự phức tạp vốn có của các mạng PPI, mức độ nhiễu của dữ liệu cao và các hiện tượng khác thường trong cấu trúc mạng.

1.2. **Một số phương pháp dự đoán tương tác Proteins điển hình**

Protein và tương tác giữa các protein là trung tâm của hầu hết các quá trình sinh học. Thông thường, protein hiếm khi hoạt động một cách độc lập mà thực hiện chức năng của chúng thông qua sự tương tác với các đơn vị phân tử sinh học khác. Do đó, việc kiểm tra các tương tác protein-protein (PPI) là cần thiết để hiểu các cơ chế phân tử của các quá trình sinh học. Trong những thập kỷ qua, nhiều kỹ thuật tiên tiến giúp phát hiện PPIs đã được phát triển như phương pháp sàng lọc 2 lai (Y2H), ái lực thanh lọc tandem (TAP), phương pháp quang phổ khối lượng để phân tích loại protein (Ms PCI) và các kỹ thuật tiên tiến khác. Một số lượng lớn các dữ liệu PPIs cho các loài khác nhau đã xây dựng. Tuy nhiên, các phương pháp thực nghiệm rất tốn kém chi phí và thời gian, vì vậy các cặp PPI thu được từ thực nghiệm chỉ chiếm một phần nhỏ các mạng tương tác Protein hoàn chỉnh. Ngoài ra, các phương pháp thực nghiệm có tỉ lệ cao các dự đoán dương tính giả và âm tính giả. Do đó, việc phát triển các phương pháp tính toán đáng tin cậy tạo thuận lợi cho việc xác định các PPIs có ý nghĩa thực tiễn to lớn.

Một số lượng lớn các phương pháp tính toán đã được đề xuất để dự đoán các tương tác Proteins.

1.2.1. **Dự đoán dựa trên thông tin các chuỗi**

Các dự đoán PPIs được thực hiện bằng cách kết hợp các thông tin của các tương tác đã biết với các thông tin liên quan đến sự tương đồng trình tự. Phương pháp này dựa trên khái niệm là một sự tương tác được tạo ra trong một loại có thể được sử dụng để dự đoán một tương tác trong một loại khác.

Yanay Ofran và Burkhard Rost [7] đã mô tả một mạng notron để xác định các các mặt tương tác từ chuỗi. Mạng notron có một lớp ẩn với 189 đơn vị đầu vào, 300 đơn vị ẩn, và hai đơn vị đầu ra (vị trí tương tác hoặc không tương tác) và sử dụng thuật toán lan truyền ngược để huấn luyện các mạng notron trên các cửa sổ làm việc (windows) của chín dư lượng liên tiếp trong chuỗi. Một cửa sổ làm việc được định nghĩa là vị trí tương tác, nếu dư lượng trung tâm tiếp xúc với một dư lượng trong một protein khác. Nhóm tác giả thực hiện huấn luyện trên 2/3 của tập dữ liệu và thử nghiệm trên 1/3 tập dữ liệu còn lại. Sau đó xoay vòng, như vậy mỗi protein sẽ được kiểm tra một lần, tức là nhóm tác giả đã huấn luyện 3 ba phiên bản khác nhau của tất cả các mạng.

Nghiên cứu của nhóm tác giả đã chỉ ra rằng các vị trí tương tác được hình thành bởi các dư lượng liên tiếp trong chuỗi. Họ đã tìm ra hơn 98% các tương tác protein-protein có ít nhất một dư lượng tương tác trong vùng chuỗi lân cận của chúng, ví dụ trong 4 dư lượng N hoặc dư lượng đoạn cuối C; 80% các tương tác protein có năm hoặc nhiều hơn các tương tác với lân cận của chúng. Khi áp dụng một ngưỡng giới hạn cho các tương tác (<= 4Å), họ đã tìm thấy ít dư lượng hơn trong chuỗi lân cận. Tuy nhiên, đa số chuỗi nine-mers vẫn chứa năm hoặc nhiều hơn các dư lượng tương tác. Kết hợp với việc quan sát dư lượng tương tác có thành phần duy nhất, nghiên cứu gợi ý rằng các vị trí tương tác được phát hiện từ chuỗi đơn.

Kết quả thu được của nghiên cứu trong hầu hết các vị trí đã dự đoán (34 của 333 proteins) 94% của các dự đoán đã được thực hiện bằng thực nghiệm thì 70% các dự đoán là đúng, và dự đoán một cách chính xác ít nhất một vị trí tương tác trong 20% của các phức hợp (complexes) (66/333).

Một nghiên cứu khác của nhóm tác giả Pitre [8] sử dụng thuật toán PIPE để dự đoán sự tương tác giữa các protein dựa trên các chuỗi polypeptide ngắn giữa các cặp protein tương tác đã biết. Thuật toán này dựa trên các tương tác đã được xác định từ trước. Giả định có 2 chuỗi protein A và B, và 2 chuỗi C và D đang tương tác. Nếu một vùng (chuỗi con) a1 trong A giống một vùng trong C, và một chuỗi b1 trong B giống một vùng trong D thì A và B cũng có thể đang tương tác qua 2 chuỗi con tương ứng là a1 và b1. Khi số lượng các cặp protein đang tương tác trong cơ sở dữ liệu chứa các chuỗi a1 và b1 tăng thì a1 và b1 là trung gian thực sự của một tương tác giữa A và B. PIPE cho thấy độ nhạy 61% khi phát hiện sự tương tác protein nấm men với 89% độ đặc hiệu và độ chính xác trung bình là 75%. Tỷ lệ này của thành công khi so sánh với các kỹ thuật sinh hóa thông dụng nhất.

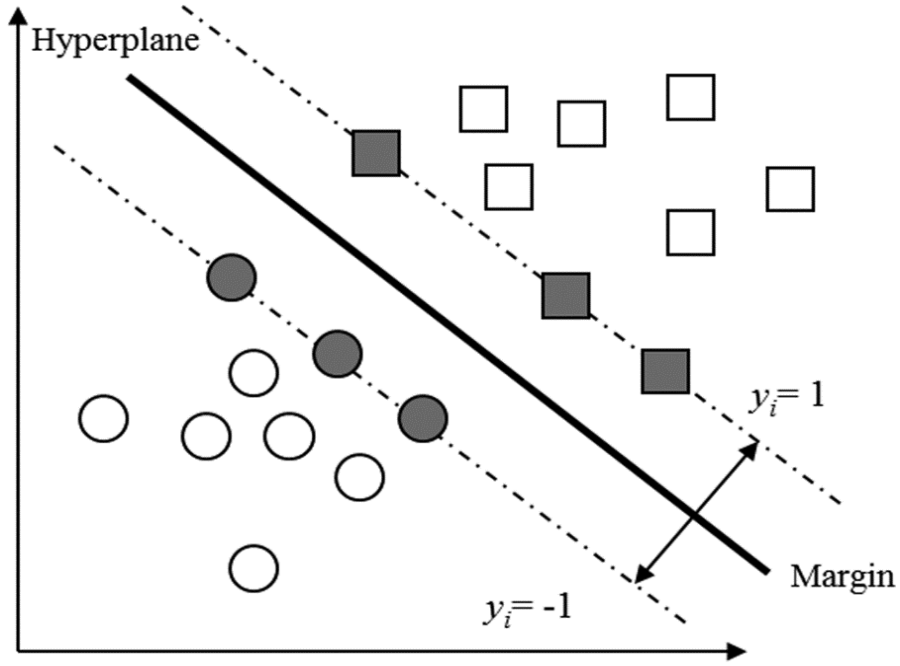
Một phương pháp dự đoán dựa trên chuỗi khác được đề xuất bởi Wojcik và Schachter [5] đếm các dữ liệu miền của các Proteins. Trong khi các tương tác thường xảy ra giữa các miền Protein, thông tin miền của mỗi protein tương tác trong một loại có thể giúp dự đoán các tương tác trong một loại khác. Trong phương pháp này, dữ liệu PPI cho một sinh vật nguồn được chuyển đổi thành một bản đồ tương tác của một nhóm miền. Các nhóm miền này được tạo bởi các miền liên kết tương tác với các vùng hoặc các miền có sự tương đồng cao về trình tự. Dữ liệu miền sau đó được xây dựng từ sự liên kết của các trình tự miền trong cùng một nhóm. Trong đó, hai nhóm miền được liên kết nếu số tương tác giữa chúng vượt quá 1 ngưỡng giới hạn. Cuối cùng, mỗi nhóm miền được ánh xạ tới một tập các protein giống nhau trong một sinh vật đích. Sự dự đoán các tương tác protein được trên sự kết nối giữa các nhóm miền.

Mẫu các miền xuất hiện trong các proteins tương tác đã biết cũng có thể giúp dự đoán các PPI khác. Sprinzak và Margalit [16] đã đề xuất sử dụng các cặp miền gọi là các chữ kí chuỗi xảy ra thường xuyên trong các proteins tương tác khác nhau. Trước tiên, các chuỗi Proteins được biểu thị bởi các chữ kí chuỗi của chúng và thu được một

bảng số liệu thống kê, sau đó các cặp chữ kí đại diện được xác định bởi việc so sánh các tần số quan sát những cá thể mà sẽ phát sinh một cách ngẫu nhiên. Phương pháp này dựa trên giả thuyết rằng tất cả các tương tác xảy ra trong các tương tác miền đã được xác định rõ.

1.2.1.1. **Mô hình dựa trên thuật toán SVM**

Máy vector hỗ trợ (Support Vector Machine - SVM) được đề xuất bởi V. Vapnik và các đồng nghiệp của ông vào những năm 1970 ở Nga, và sau đó đã trở lên nổi tiếng và phổ biến vào những năm 1990. SVM là một phương pháp phân lớp tuyến tính (Linear classifier), với mục đích xác định một siêu phẳng (hyperplane) để phân tách hai lớp của dữ liệu. Lề (margin) của siêu phẳng được xác định bằng khoảng cách giữa các mẫu dương và mẫu âm gần mặt siêu phẳng nhất. SVM sẽ lựa chọn mặt siêu phẳng phân tách có lề lớn nhất.



Hình 1.1 Phương pháp SVM

Nhóm tác giả Juwen Shen, Jian Zhang, Xiaomin Luo, Weiliang Zhu, Kunqian Yu, Kaixian Chen, Yixue Li, và Hualiang Jiang [4] đề xuất mô hình dự đoán tương tác Protein chỉ dựa trên thông tin các chuỗi. Trong nghiên cứu này, nhóm tác giả sử dụng phương pháp học máy dựa trên SVM kết hợp với hàm nhân và đặc trưng liên kết 3. Theo phương pháp này, mỗi chuỗi Protein được biểu diễn bằng một vector bao gồm các đặc trưng của các axit min và cặp tương tác protein được mô tả bởi hai vector của hai protein riêng biệt tiếp giáp nhau. Để giảm số chiều của vector và phù hợp với sự đột biến đồng nghĩa, 20 axit amin được nhóm lại trong một vài lớp theo lưỡng cực và khối lượng bên trong của các chuỗi. Phương pháp liên kết 3 tách các đặc trưng của các cặp proteins dựa trên sự phân loại các axit amin.

Để giảm vấn đề overfitting, trên 16000 cặp PPI được sử dụng để sinh các mô hình dự đoán. Phương pháp đánh giá chéo được sử dụng để tăng độ chính xác của dự đoán. Hai tham số quan trọng của SVM là C và ɣ được tối ưu với giá trị 128 và 0,25. Mô hình dự đoán PPI được sinh ra dựa trên thuật toán SVM và hàm nhân S. Để giảm sự phụ thuộc dữ liệu trên mô hình dự đoán, 5 tập huấn luyện và 5 tập kiểm tra được chuẩn bị bởi phương pháp lấy mẫu. Mỗi tập huấn luyện bao gồm 32486 cặp protein, một nửa các cặp protein được chọn ngẫu nhiên từ các cặp PPI dương tính, một nửa còn lại được chọn ngẫu nhiên từ các cặp PPI âm tính. Mỗi tập kiểm tra được tạo bởi 400 cặp protein khác. Như vậy, 5 mô hình dự đoán được sinh cho 5 tập kiểm tra. Tất cả 5 mô hình có độ chính xác lớn hơn 82,23%, độ nhạy lớn hơn 84% và độ chính xác dự đoán lớn hơn 82,75%. Mô hình có độ dự đoán chính xác trung bình 83,90±1.29%.

Phương pháp của nhóm tác giả cũng cho phép dự đoán các mạng tương tác protein. Có 3 loại mạng đã được dự đoán là mạng 1 nhân được tạo bởi 1 protein nhân tương tác với nhiều protein khác, mạng đa nhân bao gồm một vài protein nhân tương tác với các protein khác, và mạng chéo bao gồm một vài mạng đa nhân tương tác với nhau.

1.2.1.2. Mô hình dựa trên các bộ học máy cực đoan và phân tích thành phần riêng

Trong nghiên cứu này, nhóm tác giả zhu Hong You, Ying Ke Lei, Lin Zhu, Junfeng Xia và Bing Wang [6] dùng mô hình học máy cực đoan dự đoán các tương tác Protein chỉ sử dụng thông tin của các chuỗi Proteins. Trong phương pháp này, 11.188 cặp protein lấy từ cơ sở dữ liệu DIP được mã hóa thành các vector bằng cách sử dụng bốn loại protein trình tự thông tin và sử dụng phương pháp PCA (Principle component Analysis) để giảm số chiều của vector. Sau đó các học máy cực đoan được huấn luyện và tổng hợp kết quả trong một bộ phân loại. Các máy học cực đoan loại bỏ sự phụ thuộc của kết quả trên trọng lượng ngẫu nhiên ban đầu và cải thiện hiệu suất dự đoán.

Khi thực hiện trên dữ liệu PPI của nấm men Saccharomyces cerevisiae, phương pháp dự đoán chính xác 87,00% với độ nhạy 86,15% ở độ chính xác 87,59%. Một vài thí nghiệm đã được thực hiện để so sánh với phương pháp SMV (Support Vector Machine), kết quả của thí nghiệm đã chứng minh rằng, phương pháp PCA – EELM (Principle Component Analysis – Ensemble Extreme Learning Machine) thực hiện tốt hơn phương pháp SVM dựa trên phương pháp đánh giá chéo. Ngoài ra, phương pháp PCA – EELM thực hiện nhanh hơn phương pháp PCA – SVM.

1.2.2. Dự đoán dựa trên thông tin về cấu trúc protein

1.2.2.1. Mô hình PrISE.

Việc xác định các dư lượng trong các vị trí tương tác Protein có ý nghĩa quan trọng trong nhiều lĩnh đặc biệt giúp hiểu rõ các cơ chế sinh học và trong điều chế thuốc. Dựa trên việc quan sát tập các dư lượng giao diện của một loại protein có xu hướng được lưu giữ ngay cả giữa những đồng đẳng cấu trúc xa, nhóm tác giả Rafael A Jordan, Yasser EL-Manzalawy, Drena Dobbs và Vasant Honavar [11] đã giới thiệu mô hình PrISE – một phương pháp tính toán dựa trên sự tương đồng cấu trúc địa phương cho việc dự đoán các dư lượng giao diện Protein – Protein.

Nhóm tác giả đã mô tả các dư lượng bề mặt của một protein trong các phần tử cấu trúc. Mỗi phần tử này bao gồm một dư lượng chính và các dư lượng lân cận của nó. Phương pháp PrISE dùng một phần tử cấu trúc đại diện bắt thành phần nguyên tử và diện tích tiếp xúc bề mặt của các dư lượng tạo nên mỗi cấu trúc. Mỗi thành phần của PrISE xác định mỗi phần tử cấu trúc trong chuỗi Protein. Phương pháp $PrISE_L$ dựa trên sự giống nhau giữa các phần tử cấu trúc (sự tương đồng cấu trúc địa phương). Phương pháp $PrISE_G$ dựa trên sự giống nhau giữa các bề mặt Protein (sự tương đồng cấu trúc toàn cục). $PrISE_C$ kết hợp sự tương đồng cấu trúc địa phương và sự tương đồng cấu trúc toàn cục để dự đoán các dư lượng giao diện. Sự dự đoán này sẽ gán nhãn dư lượng trung tâm của phần tử cấu trúc trong chuỗi Protein như một dư lượng giao diện nếu hầu hết các phần tử cấu trúc được đánh trọng số giống với nó là các dư lượng giao diện, nếu không thì sẽ được gán nhãn dư lượng phi giao diện. Kết quả của các thí nghiệm của nhóm tác giả với việc sử dụng 3 tập dữ liệu chuẩn chỉ ra rằng phương pháp $PrISE_C$ thì nhanh hơn phương pháp $PrISE_L$ và $PrISE_G$. Nhóm tác giả cũng so sánh phương pháp $PrISE_C$ với PredUs (phương pháp dự đoán các dư lượng giao diện của chuỗi Protein dựa trên các dư lượng giao diện tương đồng về cấu trúc đã biết.) cho thấy phương pháp $PrISE_C$ có hiệu năng cao hơn hoặc tương đương PredUs khi chỉ dựa trên sự tương đồng về cấu bề mặt địa phương.

Phương pháp có sẵn ở địa chỉ: http://ailab.ist.psu.edu/software.html

1.2.2.2. Mô hình Zhang

Nhóm tác giả Qiangfeng Cliff Zhang, Donald Petrey, Lei Deng, Li Qiang, Yu Shi, Chan Aye Thu, Brygida Bisikirska, Celine Lefebvre, Domenico Accili, Tony Hunter, Tom Maniatis, Andrea Califano, và Barry Honig [9] nghiên cứu sự tương tác giữa các protein dựa trên cấu trúc trên một hệ gen sử dụng thuật toán PrePPI (Predicting protein – protein interactions) kết hợp sự tương tác phi cấu trúc và cấu trúc dùng mạng Bayesian. Thuật toán được mô tả như sau : cho 1 cặp protein QA và QB, sau đó sắp xếp chuỗi để xác định các đại diện cấu trúc MA và MB tương ứng với các mô hình tương đồng hoặc các cấu trúc được xác định bằng thực nghiệm. Sau đó sắp xếp cấu trúc để tìm lân cận NA_i và NB_j của MA và MB (trung bình mỗi cấu trúc sẽ tìm được 1500 lân cận). Bất cứ khi nào NA_i và NB_j của trên 2 triệu cặp lân cận của MA và MB tạo thành một phức hợp sẽ định nghĩa một mẫu cho mô hình tương tác của QA và QB.

Mô hình của phức hợp được tạo bởi việc đặt các cấu trúc đại diện lên trên các lân cận tương ứng trong mẫu (ví dụ MA trên NA_1, MB trên NB_2). Quy trình này sẽ tạo ra khoảng 550 triệu mô hình tương tác cho khoảng 2,4 triệu PPIs bao gồm 3900 proteins men và khoảng 12 tỉ mô hình cho khoảng 36 triệu PPIs bao gồm 13000 proteins người.

Nhóm tác giả đã tính toán 5 điểm dựa trên cấu trúc cho mỗi mô hình và sử dụng mạng Bayesian kết hợp với các điểm này để đánh giá một mô hình tương tác dựa trên tập HC và tập N. Mạng Bayesian được huấn luyện trên các tập dữ liệu chuẩn kết hợp dữ liệu tương tác từ nhiều cơ sở dữ liệu để đảm bảo mức độ bao phủ các tương tác đúng. Tập dữ liệu được chia thành 2 tập con HC và LC. Tất cả các PPIs trong bộ gen đã cho không nằm trong tập HC và LC tạo thành tập N. Dùng phân lớp Bayesian huấn luyện trên tập HC nhóm tác giả chọn được mô hình tương tác tốt nhất cho mỗi PPI.

Nghiên cứu sử dụng phương pháp đánh giá chéo (10 – fold cross validation). Nhóm tác giả chia các tập âm tính và dương tính thành 10 tập con có cùng kích thước, mỗi lần sử dụng 9 tập con để huấn luyện phân lớp, 1 tập con sử dụng để kiểm tra và lặp lại 10 lần trên các tập con khác nhau. Nhóm tác giả đếm số dương tính thật (dự đoán trong tập HC) và dương tính giả (dự đoán trong tập N) và tính toán tỉ lệ dương tính thật = TP/(TP+FN) và tỉ lệ dương tính giả = FP/(FP+TN).

1.2.2.3. **Mô hình iLoops**

Trong nghiên cứu này, nhóm tác giả Joan Planas-Iglesias, Manuel A. Marin-Lopez, Jaume Bonet, Javier Garcia- Garcia và Baldo Oliva [10] trình bày về iLoops, một máy chủ web có thể dự đoán các tương tác protein dựa trên đặc điểm cấu trúc địa phương.

Đầu vào của các máy chủ iLoops là: trình tự của các protein truy vấn và các cặp proteins cần kiểm tra. Các cặp đặc điểm cấu trúc (được tạo bởi các vòng hoặc các miền) được phân loại theo khả năng kích thích hoặc ức chế một tương tác protein-protein, tùy thuộc vào sự quan sát các cặp proteins tương tác và không tương tác. Máy chủ đánh giá sự tương tác sử dụng bộ phân loại tập hợp ngẫu nhiên.

Nghiên cứu sử dụng phân loại các vòng từ ArchDB, và các miền từ Scop để xác định các đặc điểm cấu trúc địa phương (SFs). SFs là các miền hoặc các vòng. Chữ ký protein được định nghĩa là nhóm 1-3 SFs (có thể là vòng hoặc miền). Chữ ký tương tác được định nghĩa là các cặp protein - chữ ký cùng kiểu giữa hai loại protein tương tác hoặc không-tương tác.

Nhóm tác giả sử dụng mô hình RF để dự đoán. RF là mô hình được sinh từ WEKA bằng cách học các chữ ký của tập huấn luyện. Máy chủ sử dụng các mô hình RF để dự đoán các cặp giá trị đầu vào.

Server iLoops có sẵn ở địa chỉ : http://sbi.imim.es/iLoopsServer/. Dữ liệu đầu vào máy chủ web iLoops được đưa vào qua một vùng văn bản và người sử dụng chọn các loại SFs cho sự dự đoán. Mỗi lần kiểm tra được tối đa 25 cặp protein. Các máy chủ cung cấp một mã nhận dạng lấy nhận các dự đoán. Các dự đoán có thể được duyệt thông qua giao diện web hoặc tải về trong một tập tin xml. Sự chính xác dự đoán phụ thuộc vào giá trị UR (tỉ lệ mất cân bằng) giữa PPIs và NIPs. Giá trị UR này được chọn bởi người sử dụng và phụ thuộc vào các điều kiện thí nghiệm (ví dụ đối với bất kỳ cặp protein trên người thì giá trị UR xấp xỉ 1/50). Theo mặc định máy chủ sẽ chọn giá trị RC (mức độ giảm tỉ lệ dự đoán dương tính giả) tốt nhất cho UR. Tùy chọn nâng cao cho phép người dùng chọn các giá trị RCs khác nhau cho bộ phân loại RF.

Máy chủ iLoops cung cấp một giao diện thân thiện với người dùng. Các kết quả dự đoán có thể được truy vết lại cơ sở dữ liệu ban đầu để hiểu rõ các kết quả dự đoán. Máy chủ iLoops cũng cung cấp khả năng chọn lựa tỷ lệ dự kiến giữa PPI và NIPS trong việc dự đoán. Như vậy, các máy chủ cho phép người dùng giải quyết truy vấn khác nhau như dự đoán số tương tác thực lớn nhất hoặc giảm thiểu các dự đoán sai. Sự dự đoán trong mỗi tập cân bằng (1 PPI cho mỗi 1 NTP) có thể đạt độ chính xác 89% và hồi tưởng là 81%. Sự dự đoán được thực hiện trên một tập chứa 1 PPI cho 50 NIPs có thể thể đạt độ chính xác là 38% và độ hồi tưởng là 39%.

CHƯƠNG 2. KỸ THUẬT HỌC SÂU (DEEP LEARNING)

2.1. Giới thiệu về học sâu (Deep Learning)

Từ năm 2006, học kiến trúc sâu hay thường gọi là học sâu đã nổi lên như một lĩnh vực nghiên cứu mới của học máy [29,30]. Trong những năm qua, các kỹ thuật được phát triển từ sự nghiên cứu học sâu đã ảnh hưởng tới một loạt các lĩnh vực quan trọng của học máy và trí tuệ nhân tạo. Trước tiên, ta sẽ tìm hiểu một vài định nghĩa về học sâu:

- Học sâu là một lớp của các kỹ thuật học máy mà khai thác nhiều lớp của sự xử lý thông tin phi tuyến tính cho sự biến đổi và trích đặc trưng giám sát hoặc không giám sát và cho việc phân tích và phân loại mẫu (Li Deng et al., 2014, page 10).
- Học sâu là một lớp của các thuật toán học máy mà sử dụng nhiều lớp các đơn vị xử lý phi tuyến tính cho sự biến đổi và trích đặc trưng. Mỗi lớp sẽ sử dụng đầu ra của lớp trước đó như là giá trị đầu vào. Các thuật toán này có thể là học giám sát hoặc không giám sát. Các ứng dụng bao gồm phân tích mẫu (không giám sát) và phân loại mẫu (có giám sát). Học sâu dựa trên việc học đa lớp các đặc trưng hoặc sự biểu diễn của dữ liệu. Trong đó, các đặc trưng cấp cao hơn thu được từ các đặc trưng cấp thấp hơn để tạo thành một biểu diễn theo thứ bậc (Wikipedia on Deep learning).

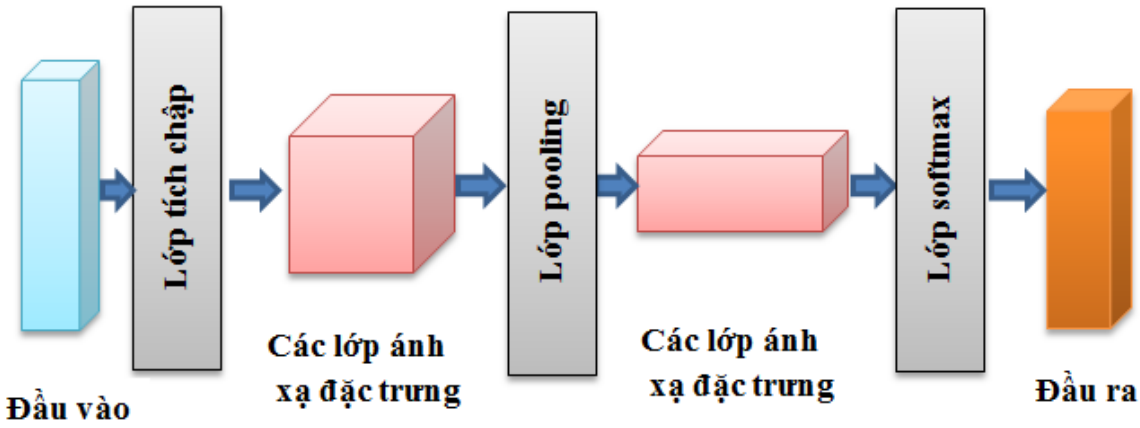
2.2. Phân loại mạng học sâu (Deep Learning)

Học sâu ám chỉ một lớp khá rộng các kiến trúc và các kỹ thuật học máy sử dụng nhiều lớp xử lý thông tin phi tuyến có tính phân cấp. Tùy thuộc vào cách các kiến trúc và các kỹ thuật được sử dụng người ta có thể phân loại các công việc trong lĩnh vực này thành ba nhóm chính:

- Các mạng sâu cho học không giám sát: nhằm đạt được mối quan hệ bậc cao của dữ liệu được quan sát cho mục đích phân tích hoặc tổng hợp mẫu khi thông tin về các nhãn lớp không có sẵn.
- Các mạng sâu cho học có giám sát: cung cấp khả năng phân loại cho mục đích phân loại mẫu. Dữ liệu nhãn đích luôn luôn có sẵn dưới các hình thức trực tiếp hoặc gián tiếp cho học có giám sát.
- Các mạng sâu lai: mục đích phân loại được hỗ trợ một cách đáng kể như mạng sâu không giám sát, và có thể được thực hiện một cách chuẩn hóa và tối ưu hơn các mạng sâu cùng loại. Mục tiêu cũng có thể được thực hiện khi điều kiện phân loại cho học có giám sát được sử dụng để đánh giá các tham số trong các mạng sâu không giám sát.

2.3. Mạng nơ ron tích chập (Convolutional neural network - CNN)

Mạng nơ ron tích chập [28] được hai nhà khoa học Yann LeCun và Yoshua Bengio đề xuất vào năm 1998. Cấu trúc cơ bản của một mạng nơ ron tích chập gồm bốn lớp: đầu vào, lớp tích chập, lớp pooling và đầu ra.



Hình 2.10. Cấu trúc mạng nơ ron tích chập

Trong đó, đầu vào là dữ liệu nhiều chiều. Trong luận văn thì đầu vào là chuỗi peptit được biểu diễn dưới dạng ma trận như sau:

Véc tơ của chuỗi peptit:

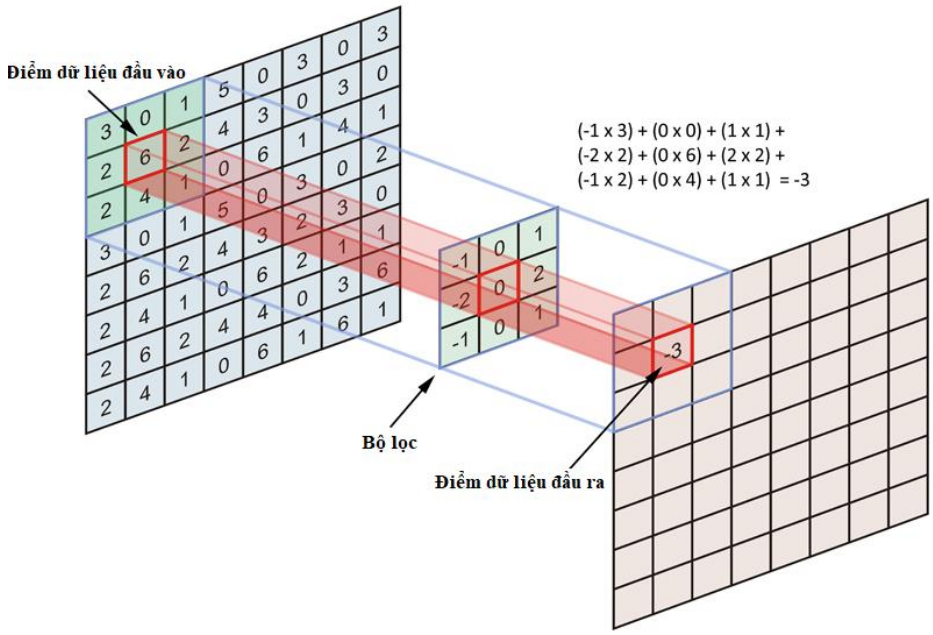
$[X_{11}, X_{12}, ..., X_{1n}, X_{21}, X_{22}, ..., X_{2n}, X_{31}, X_{32}, ..., X_{3n}, ..., X_{91}, X_{92}, ..., X_{9n}]$

Ma trận tương ứng:

$$\begin{bmatrix} X_{11}, X_{12}, ..., X_{1n} \\ X_{21}, X_{22}, ..., X_{2n} \\ X_{31}, X_{32}, ..., X_{3n} \\ ... \\ X_{91}, X_{92}, ..., X_{9n} \end{bmatrix}$$

Hình 2.11. Ma trận đầu vào của mạng CNN

Lớp tích chập là lớp đầu tiên trong mạng CNN. Thay vì kết nối tới tất cả các điểm dữ liệu đầu vào. Lớp tích chập sử dụng một bộ lọc có kích thước nhỏ (thường là 3x3 hoặc 5x5) chiếu vào một vùng dữ liệu đầu vào và tiến hành tính tích chập giữa các giá trị trên bộ lọc và giá trị trên vùng dữ liệu đầu vào được chiếu như hình dưới.



Hình 2.12. Tích chập giữa bộ lọc và vùng dữ liệu vào

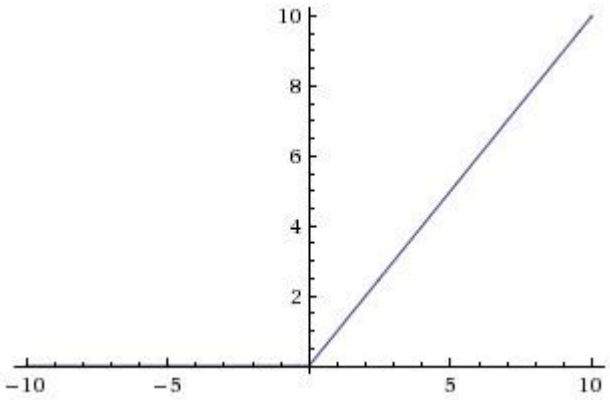
Một bộ lọc cũng được gọi là một nơ ron hoặc một kernel. Các giá trị trên bộ lọc là các trọng số hoặc tham số. Vùng dữ liệu trên dữ liệu đầu vào mà bộ lọc chiếu qua gọi là vùng tiếp nhận (receptive field). Bộ lọc sẽ lần lượt dịch chuyển và quét trên toàn bộ dữ liệu đầu vào theo một giá trị gọi là bước trượt (stride). Với mỗi lần trượt và tính tính chập sẽ thu được một giá trị, các giá trị thu được sau khi bộ lọc quét và tính tích chập gọi là ánh xạ đặc trưng (feature map). Một lớp ánh xạ đặc trưng là đầu ra của một bộ lọc áp dụng tới lớp trước đó.

Hàm kích hoạt ReLU

Hàm Rectified linear unit (ReLU) có công thức như sau:

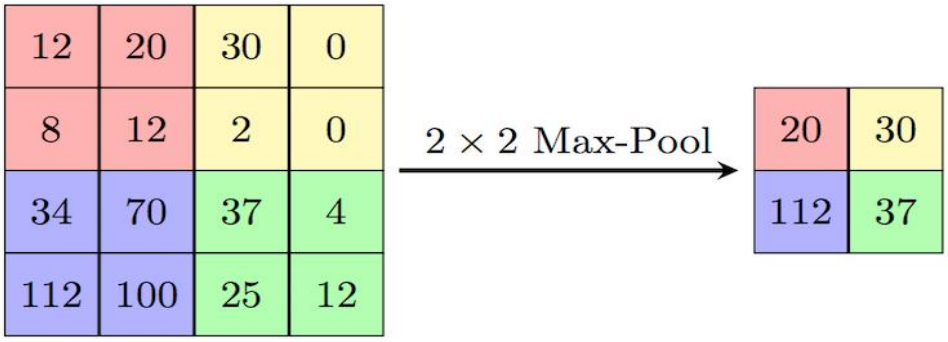
$$y = \max(0,x)$$

Hàm ReLU thường được sử dụng phía sau lớp tích chập để chuyển các kết quả âm từ lớp tích chập thành các giá trị 0. Đồ thị của hàm ReLU:



Hình 2.13. Đồ thị hàm ReLU

Lớp pooling thường theo sau một hoặc nhiều lớp tích chập. Lớp này sử dụng một bộ lọc dịch chuyển quét toàn bộ dữ liệu vào, mỗi lần dịch chuyển theo một bước trượt cho trước giống lớp tích chập nhưng lớp pooling không tính tích chập mà sẽ tiến hành lấy mẫu. Trong quá trình trượt, các giá trị đại diện cho dữ liệu vào trên vùng được trượt (vùng lấy mẫu) sẽ được giữ lại. Một số phương pháp lấy mẫu phổ biến là MaxPooling (lấy giá trị lớn nhất), MinPooling (lấy giá trị nhỏ nhất) và AveragePooling (lấy giá trị trung bình).

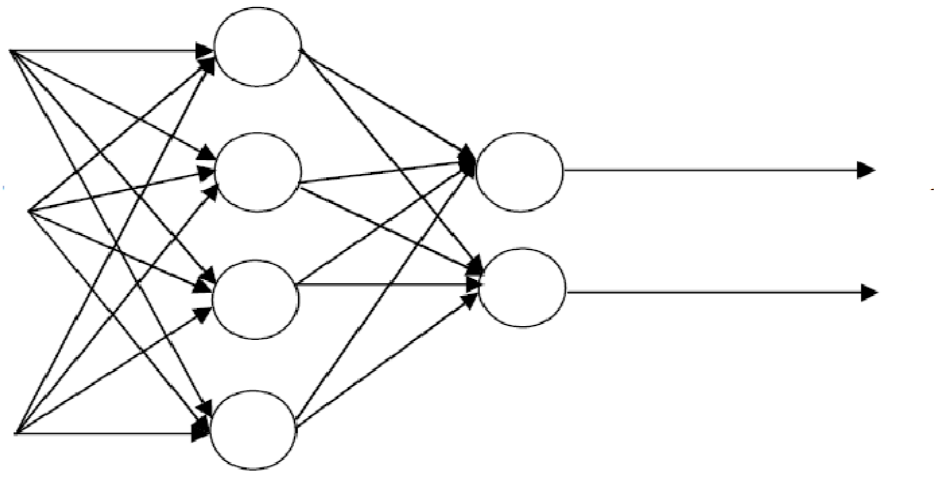


Hình 2.14. Phương pháp MaxPooling với cửa sổ 2x2 và bước trượt 2

Lớp Pooling có vai trò làm giảm kích thước dữ liệu của lớp trước đó. Với các dữ liệu có kích thước lớn khi qua lớp Pooling sẽ được giảm xuống nhưng vẫn giữ được các đặc trưng của dữ liệu. Việc giảm kích thước dữ liệu giúp giảm các tham số, tăng hiệu năng tính toán và kiểm soát hiện tượng overfitting trong quá trình huấn luyện.

Lớp kết nối đầy đủ

Lớp này được sử dụng ở cuối của mạng sau quá trình xử lý và trích chọn các đặc trưng đã được thực hiện ở các lớp tích chập và pooling. Lớp kết nối đầy đủ có cấu trúc giống như các lớp trong mạng nơ ron truyền thẳng truyền thống. Trong đó, các nơ ron trên mỗi lớp sẽ liên kết đầy đủ tới các nơ ron trên các lớp tiếp theo như hình sau:



Hình 2.15. Lớp liên kết đầy đủ

Lớp này sử dụng hàm kích hoạt Softmax để phân lớp các giá trị ánh xạ đặc trưng vào các lớp đầu ra cụ thể.

Hàm Softmax có công thức như sau:

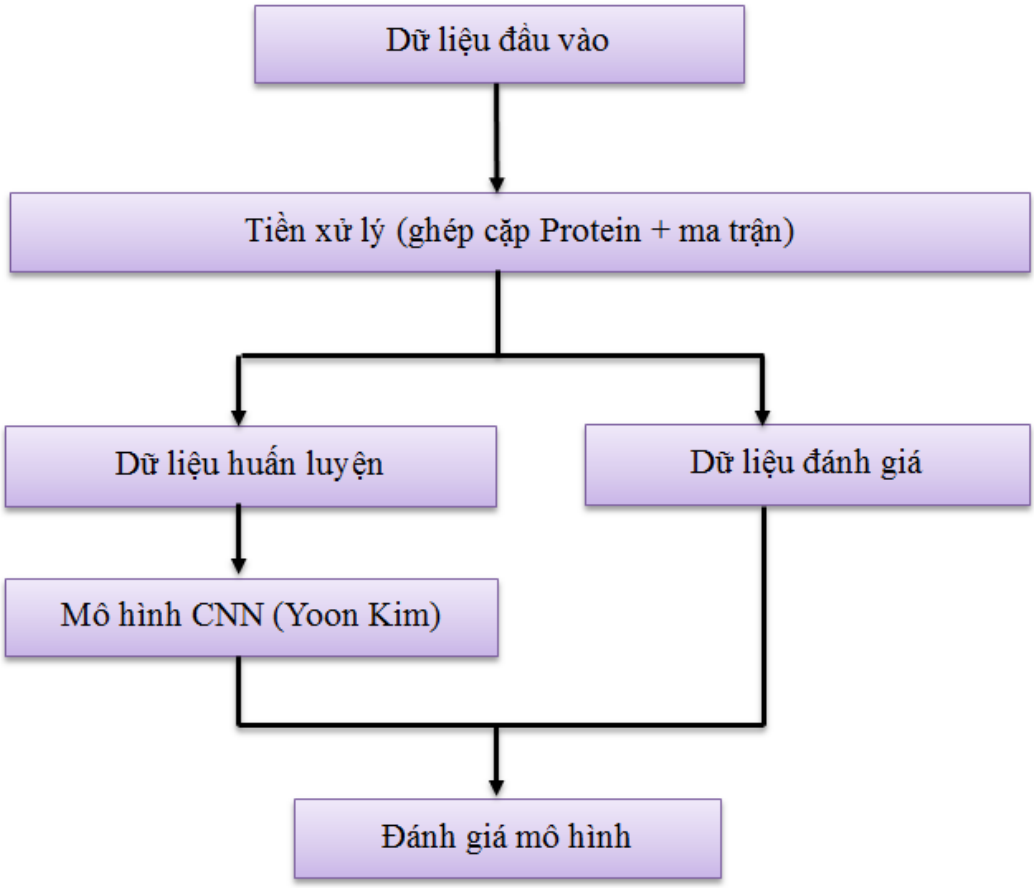
$$y_i = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}, \forall i \in [1, 2, \dots, n]$$

Hàm softmax sẽ chuyển một vector x có các giá trị bất kì về vector y chứa các giá trị dưới dạng xác suất, trong đó x = {x_1, x_2, …, x_n}, y = {y_1, y_2, …, y_n) và n là số phân lớp. Giá trị thứ i của véc tơ y kí hiệu là y_i đại diện cho xác suất để dữ liệu này thuộc vào lớp thứ i.

CHƯƠNG 3. MÔ HÌNH DỰ ĐOÁN TƯƠNG TÁC PROTEINS DỰA TRÊN KỸ THUẬT HỌC SÂU (DEEP LEARNING).

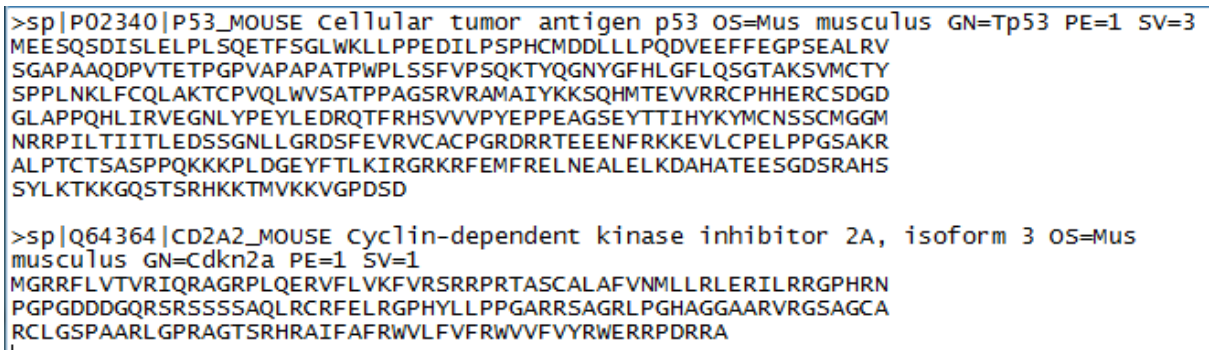
3.1. Giới thiệu về mô hình.

Phần này luận văn sẽ trình bày quá trình xây dựng mô hình dự đoán tương tác protein dựa trên kỹ thuật học sâu. Ban đầu dữ liệu đầu vào sẽ được tiến hành tiền xử lý, đó là quá trình ghép cặp và ma trận hóa dữ liệu. Sau đó, dữ liệu sẽ được chia thành hai tập riêng biệt, bao gồm tập huấn luyện và tập đánh giá. Tập huấn luyện sẽ được dùng để xây dựng mô hình. Mô hình sau khi đã xây dựng xong, tập đánh giá sẽ được đưa vào mô hình để đánh giá chất lượng mô hình. Quá trình dự đoán tương tác Proteins trong luận văn được thực hiện theo các bước sau:



Hình 3.1. Quá trình dự đoán tương tác proteins

Trong đó, dữ liệu đầu vào là chuỗi các amino axit trong các cặp protein tương tác và không tương tác như hình sau:



Hình 3.2. Ví dụ cặp protein tương tác

Các chuỗi amino axit này sẽ được biểu diễn dựa trên các thuộc tính lý hóa sinh. Các amino axit có các thuộc tính hóa học như tính axit, bazơ,….hay các thuộc tính vật lý như: độ tan, độ sôi,… các thuộc tính lý-hoá-sinh này sẽ được biểu diễn dưới dạng vector. Ví dụ được mô tả theo bảng sau:

Bảng 3.1 [1]Cách tính véc tơ của amino axit

Amino axit	**Thuộc tính**				**Véc tơ**
	1	**2**	**…**	**544**	
$\mathbf{X_1}$	$\mathbf{X_{1\text{-}1}}$	$\mathbf{X_{1\text{-}2}}$		$\mathbf{X_{1\text{-}544}}$	$[\mathbf{X_{1\text{-}1}}, \mathbf{X_{1\text{-}2}}, \ldots, \mathbf{X_{1\text{-}544}}]$
$\mathbf{X_2}$	$\mathbf{X_{2\text{-}1}}$	$\mathbf{X_{2\text{-}2}}$		$\mathbf{X_{2\text{-}544}}$	$[\mathbf{X_{2\text{-}1}}, \mathbf{X_{2\text{-}2}}, \ldots, \mathbf{X_{2\text{-}544}}]$
…	…	…	…	…	…
$\mathbf{X_{20}}$	$\mathbf{X_{20\text{-}1}}$	$\mathbf{X_{20\text{-}2}}$		$\mathbf{X_{20\text{-}544}}$	$[\mathbf{X_{20\text{-}1}}, \mathbf{X_{20\text{-}2}}, \ldots, \mathbf{X_{20\text{-}544}}]$

Tập các thuộc tính này được lấy từ cơ sở dữ liệu AAIndex. AAIndex [30] là cơ sở dữ liệu các thuộc tính lý – hoá - sinh, bao gồm ba tập dữ liệu: AAIndex1, AAIndex2 và AAIndex3. Luận văn này sẽ sử dụng dữ liệu từ tập AAIndex1 với 544 thuộc tính. Một protein có tối đa 20 loại amino axit. Như vậy mỗi amino axit sẽ là một véc tơ 544 chiều.

Quá trình tiền xử lý dữ liệu vào sẽ được tiến hành bằng cách ghép cặp protein. Protein P_1 và protein P_2 sẽ được ghép thành cặp P_1P_2.

Chuỗi protein P_1 có dạng:

$$P_1 = A_{11}A_{12}...A_{1n}$$

trong đó, A_{1i} (i=1..n) là amino axit trong 20 loại amino axit.

Chuỗi protein P_2 có dạng:

$$P_2 = A_{21}A_{22}...A_{2m}$$

trong đó, A_{2j} (j=1..m) cũng là amino axit trong 20 loại amino axit.

[1] Trác Quang Thịnh (2017), Nghiên cứu so sánh các phương pháp biểu diễn chuỗi peptit trong bài toán dự đoán vị trí protein bị phốt pho hóa, ĐHQGHN.

Như vậy, với mỗi cặp protein (P_1, P_2) sẽ tạo thành một chuỗi có dạng :

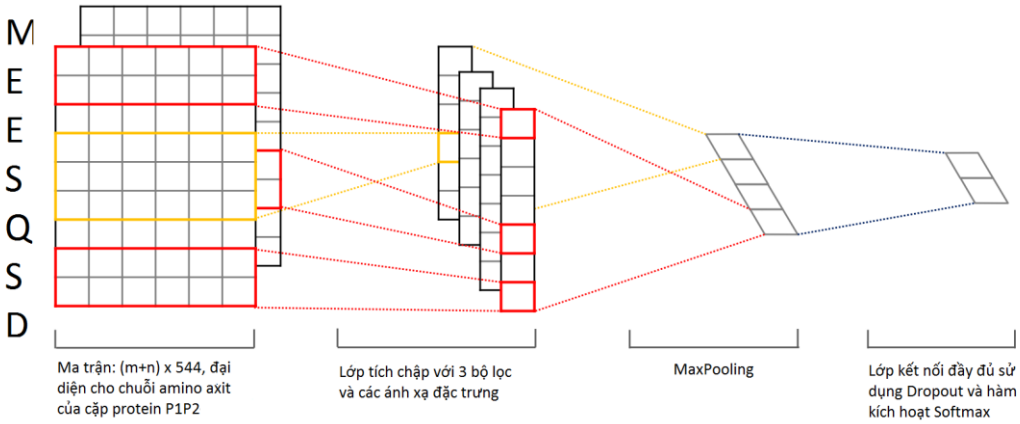
$$(P_1, P_2) = A_{11}A_{12}...A_{1n}\ A_{21}A_{22}...A_{2m}$$

Với mỗi amino axit A_{1i} (i=1..n) và A_{2j} (j=1..m) sẽ có một vector 544 chiều. Như vậy, cặp (P_1, P_2) tạo ra một ma trận có kích thước (n+m)*544.

Dữ liệu sau khi được tiền xử lý sẽ được đưa vào mô hình để huấn luyện.

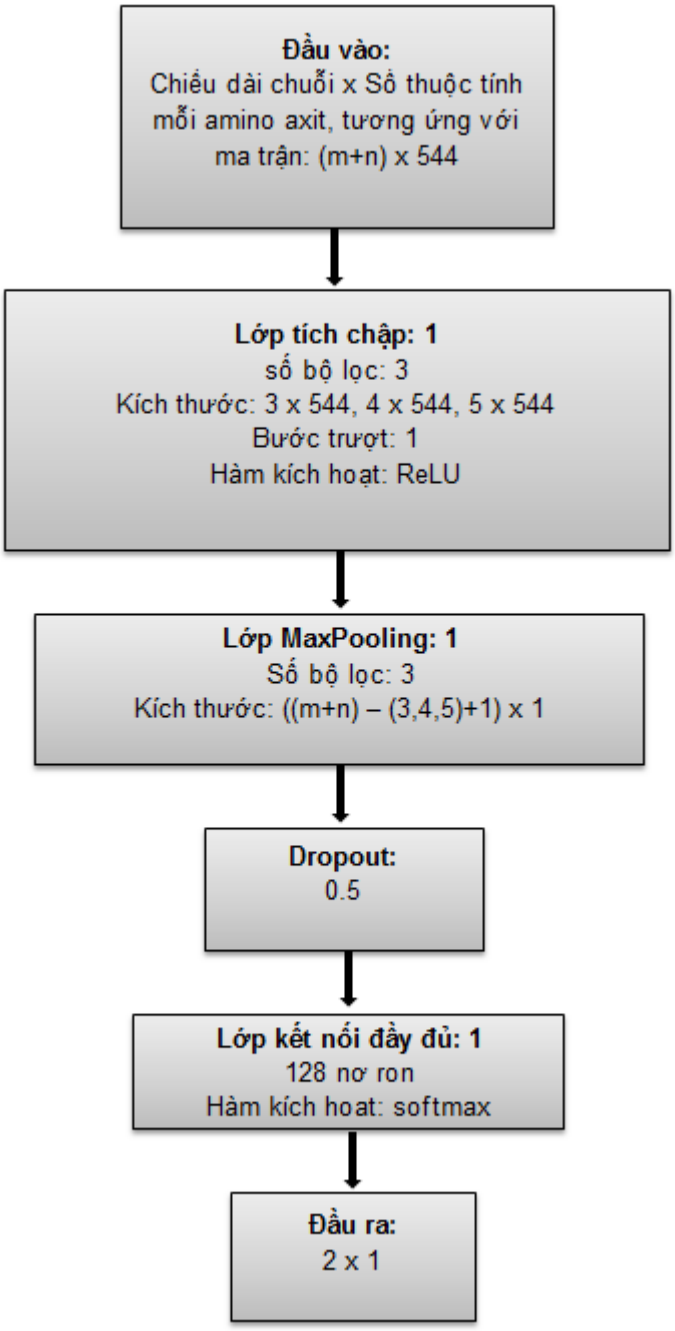
3.2. Xây dựng mô hình

Luận văn sử dụng mô hình của Yoon Kim [29] trong bài toán phân loại câu. Mô hình có dạng như sau:



Hình 3.3. Mô hình dự đoán tương tác Protein

Cụ thể mô hình được xây dựng như sau:



Hình 3.4. Mô hình dự đoán với các thông số cụ thể

Mô hình gồm các lớp như sau:

Lớp đầu vào là ma trận có kích thước ((m+n) x 544) tương ứng với chiều dài của hai chuỗi amino axit của hai protein được ghép cặp và 544 là số thuộc tính lý hóa sinh của mỗi amino axit.

Một lớp tích chập sử dụng ba bộ lọc có kích thước (3 x 544), (4 x 544), (5 x 544), với bước trượt bằng 1 và sử dụng hàm kích hoạt ReLU.

Sau đó là một lớp maxpooling với 3 bộ lọc có kích thước ((m+n -4) x 1), ((m+n -5) x 1), ((m+n -6) x 1), sử dụng dropout là 0.5 (dropout là kỹ thuật giảm overfitting).

Mạng sử dụng một lớp kết nối đầy đủ với 128 nơ ron và sử dụng hàm softmax để phân lớp đầu ra. Trong đó, đầu ra gồm 2 giá trị (10 là không tương tác, 01 là tương tác).

3.3. Nguồn dữ liệu tương tác giữa các protein

Luận văn sử dụng cơ sở dữ liệu DIP [32] là cơ sở dữ liệu chứa các cặp protein tương tác và cơ sở dữ liệu Negatome [33] chứa các cặp protein không tương tác. Mỗi cơ sở dữ liệu này chứa 6445 cặp Protein. Đây là 2 cơ sở dữ liệu được sử dụng phổ biến trong các nghiên cứu về Protein và vẫn thường xuyên được cập nhật.

3.4. Đánh giá mô hình

Luận văn sử dụng phương pháp đánh giá chéo (k-fold cross validation) với k = 5 để đánh giá mô hình. Dữ liệu đầu vào sẽ được chia thành 5 phần có tỉ lệ dữ liệu dương / dữ liệu âm bằng nhau trên tất cả các phần. Sau đó, phương pháp này sẽ thực hiện một vòng gồm k = 5 lần lặp, tại mỗi lần lặp, 1 phần dữ trên tổng số 5 phần dữ liệu sẽ làm đầu vào để xây dựng mô hình, phần dữ liệu còn lại dùng để đánh giá chất lượng mô hình. Để đảm bảo việc đánh giá mang tính chính xác thì phần dữ liệu dùng để xây dựng mô hình không chứa bất kì phần tử nào của phần dữ liệu dùng để đánh giá.

Trong luận văn, ma trận nhầm lẫn cũng được sử dụng để đánh giá chất lượng mô hình:

Bảng 3.2. Ma trận nhầm lẫn

Lớp c		Dự đoán	
		Thuộc	Không thuộc
Kết quả thực	Thuộc	TP	FN
	Không thuộc	FP	TN

trong đó TP là số các trường hợp thuộc lớp c được dự đoán đúng, FP là số các trường hợp không thuộc lớp c bị dự đoán nhầm vào lớp c, FN là số các trường hợp thuộc lớp c bị dự đoán nhầm không thuộc lớp c là TN là số các trường hợp không lớp c được dự đoán đúng.

Luận văn cũng sử dụng AUC (diện tích dưới đường cong) [31] làm độ đo để đánh giá chất lượng mô hình.

Cụ thể với số lượng cặp Protein huấn luyện (Pos/Neg) là 1000/1000, kết quả đạt được cụ thể như sau:

Bảng 3.3 thống kê các độ đo mô hình trên tập huấn luyện

Độ đo Recall	0.852
Độ đo Precision	0.845
Độ đo F1	0.845

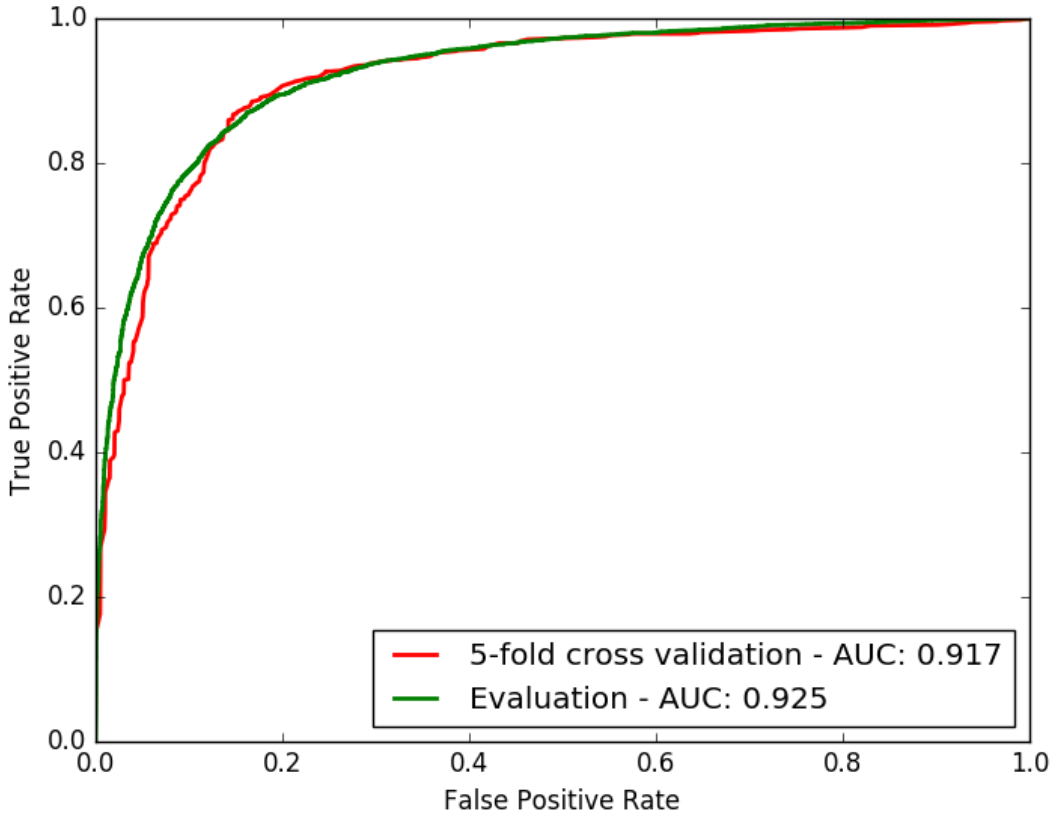
Độ chính xác Accuracy	0.846
AUC	0.917

Sau khi xây dựng được mô hình, Số lượng cặp Protein (Pos/Neg) dùng để đánh giá là 5445/5445, kết quả đạt được như sau:

Bảng 3.4 thống kê các độ đo mô hình dự đoán trên tập đánh giá

Độ đo Recall	0.788
Độ đo Precision	0.888
Độ đo F1	0.835
Độ chính xác Accuracy	0.844
AUC	0.925

Độ đo AUC được thể hiện qua đường cong ROC như sau:



Hình 3.5. Đồ thị thể hiện độ đo AUC

KẾT LUẬN

Kết quả đạt được

Luận văn đã nghiên cứu tổng quan về tương tác protein và bài toán dự đoán tương tác protein cũng như khái quát các kiến thức cơ bản của kỹ thuật học sâu, và đi sâu vào nghiên cứu mạng nơ ron tích chập. Đồng thời xây dựng thành công mô hình dự đoán tương tác protein sử dụng mạng nơ ron tích chập. Mô hình được xây dựng từ 3000 cặp protein tương tác và 3000 cặp protein không tương tác. Mô hình được đánh giá thông qua phép kiểm định chéo với k = 10 và sử dụng ma trận nhầm lẫn, độ đo AUC để đánh giá chất lượng mô hình. Mô hình dự đoán đạt được kết quả tương đối tốt với độ chính xác 0.89.

Hướng phát triển

Với những kết quả đã đạt được, luận văn sẽ tiếp tục nghiên cứu để tăng độ chính xác chất lượng mô hình dự đoán thông qua việc tăng số lượng cặp protein đưa vào huấn luyện, cũng như tìm cách tối ưu các tham số trong mô hình, đồng thời luận văn sẽ tiếp tục nghiên cứu các phương pháp tiên tiến khác được đề xuất gần đây để so sánh đánh giá trên bài toán dự đoán tương tác protein giúp có cái nhìn sâu sắc hơn về phương pháp học sâu.

TÀI LIỆU THAM KHẢO

Tiếng Việt

[1]. Nguyễn Văn Vinh (2015), *Slides bài giảng trí tuệ nhân tạo nâng cao*, ĐH Quốc gia Hà Nội, Hà Nội.

[2]. Phan Xuân Hiếu (2016), *Slides bài giảng Khai phá dữ liệu*, Đại học Quốc gia Hà Nội, Hà Nội.

[3]. Nguyễn Văn Cách (2005), *Tin sinh học*, Nhà xuất bản Khoa học và kỹ thuật, Hà Nội.

Tiếng Anh

[4]. Juwen Shen, Jian Zhang, Xiaomin Luo, Weiliang Zhu, Kunqian Yu, Kaixian Chen, Yixue Li, Hualiang Jiang (2006), "Predicting protein – protein interactions based only on sequences information", *PNAS*, 104 (11): 4337 – 4341.

[5]. Wojcik, J. and Schachter (2001), "Protein–protein interaction map inference using interact ing domain profile pairs", *Bioinformatics*, 17:S296–S305.

[6] Zhu-Hong You, Ying-Ke Lei, Lin Zhu, Junfeng Xia, Bing Wang (2013), "Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis", *BMC Bioinformatics*, 14(Suppl 8): S10.

[7]. Yanay Ofrana, Burkhard Rosta (2003), "Predicted protein-protein interaction sites from local sequence information", *FEBS Letters*, 544 236-239 FEBS 27273.

[8]. Sylvain Pitre (2006), "PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs", *BMC Bioinformatics*, 7:365 doi:10.1186/1471-2105-7-365.

[9]. Qiangfeng Cliff Zhang (2012) , "Structure-based prediction of protein-protein interactions on a genome-wide scale", *Nature*, 490(7421): 556–560. doi:10.1038/nature11503.

[10]. Joan Planas-Iglesias (2013), "iLoops: a protein-protein interaction prediction server based on structural features", *Bioinformatic, 29(18):2360-2*.

[11]. Rafael A Jordan, Yasser EL-Manzalawy, Drena Dobbs, Vasant Honavar (2012), "Predicting protein-protein interface residues using local surface structural similarity", *BMC Bioinformatics*, 10.1186/1471-2105-13-41.

[12]. Tristan T Aumentado-Armstrong, Bogdan Istrate, Robert A Murgita (2015), "Algorithmic approaches to protein-protein interaction site prediction. Algorithms for Molecular Biology", *BioMed Central*, 10:7 .

[13]. Joao P. G. L. M. Rodrigues and Alexandre M. J. J. Bonvin (2014), "Integrative computational modeling of protein interactions", *FEBS*, 1988–2003.

[14]. Aidong Zhang (2009), *Protein interaction networks*, Cambridge University Press.

[15]. Rob Brazas (2011), *In vitro and in vivo methods to study protein:protein interactions*, Promega.

[16]. Sprinzak, E. and Margalit (2001), "Correlated sequence-signatures as markers of protein - protein interaction", *Molecular Biology*, 311:681–692.

[17]. Li Deng and Dong Yu (2014), *Deep Learning: Methods and Applications*, Foundation and trends in signal processing, Volume 7 Issue 3-4, ISSN: 1932-8346.

[18]. Russ Salakutdinov (2009), *Deep Learning*, University of Toronto, Canada.

[19]. http://deeplearning.net/.

[20]. http://www.deeplearningbook.org/.

[21]. Eric Roberts (2000), Neural Networks. https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/index.html /.

[22]. Dr. G.P.Rameshkumar, S. Samundeswari (2014), *Neural Network, Artificial Neural Network (ANN) and Biological Neural Network (BNN) in Soft Computing*, Volume 30; 3(3): 1159–1163, ISSN: 2277-9655.

[23]. O.S. Eluyode and Dipo Theophilus Akomolafe (2013), "Comparative study of biological and artificial neural networks", *European Journal of Applied Engineering and Scientific Research*, 2 (1):36-46.

[24]. Warren S. Mcculloch and Walter Pitts (1943), "A logical calculus of the ideas immanent in nervous activity", *Ulletin of mathematical biophysics*, Volume 5.

[25]. Martin T. Hagan, Howard B. Demuth, Mark Hudson Beale and Orlando De Jesús (2014), *Neural Network Design* 2nd Edition.

[26]. Jeff Heaton (2008), *Introduction to Neural Networks*, Heaton Research.

[27]. Kenvil L, Priddy and Paul E .Keller (2005), *Artifical neural networks an introduction*, The international Society for Optical Engineering.

[28]. LeCun, Yann (1998), "Gradient-based learning applied to document recognition", *IEEE*, 86.11: 2278-2324.

[29]. Yoon Kim (2014), "Convolution neural networks for sentence classification", *arXiv* 1408.5882.

[30]. Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanehisa (2008), "AAindex: amino acid index database", *Nucleic Acids Res*, 28(1): 374.

[31]. DeLong ER, DeLong DM, Clarke-Pearson DL (1988), "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach", *Biometrics,* 44(3):837–845.

[32]. Salwinski L, Miller C S, Smith A J (2004), "The database of interacting proteins", *Nucleic acids research*, 32(suppl 1): D449-D451.

[33]. Smialowski P, Pagel P, Wong P (2010), "The Negatome database: a reference set of non-interacting protein pairs", *Nucleic acids research*, 38(suppl 1): D540-D544.

[34]. Tanlin Sun, Bo Zhou, Luhua Lai (2017), "Sequence-based prediction of protein protein interaction using a deep-learning algorithm", *BMC Bioinformatics*, 10.1186/s12859-017-1700-2.