Ứng dụng nhóm mô hình Graph-based
Xây dựng bài toán dự đoán bệnh dựa trên triệu chứng

Nguyễn Anh Minh

Nội dung trình bày

01
Giới thiệu bài toán

- Giới thiệu bài toán dự đoán bệnh
- Phạm vi, nội dung, mục tiêu nghiên cứu

02
Dữ liệu sử dụng

Khai phá dữ liệu

03
Mô hình nghiên cứu

- Mô hình đề xuất
- Chi tiết mô hình

04
Kết quả chạy mô hình

05
Kết luận

- Kết quả đạt được và hạn chế của đề tài.
- Hướng phát triển đề tài

01

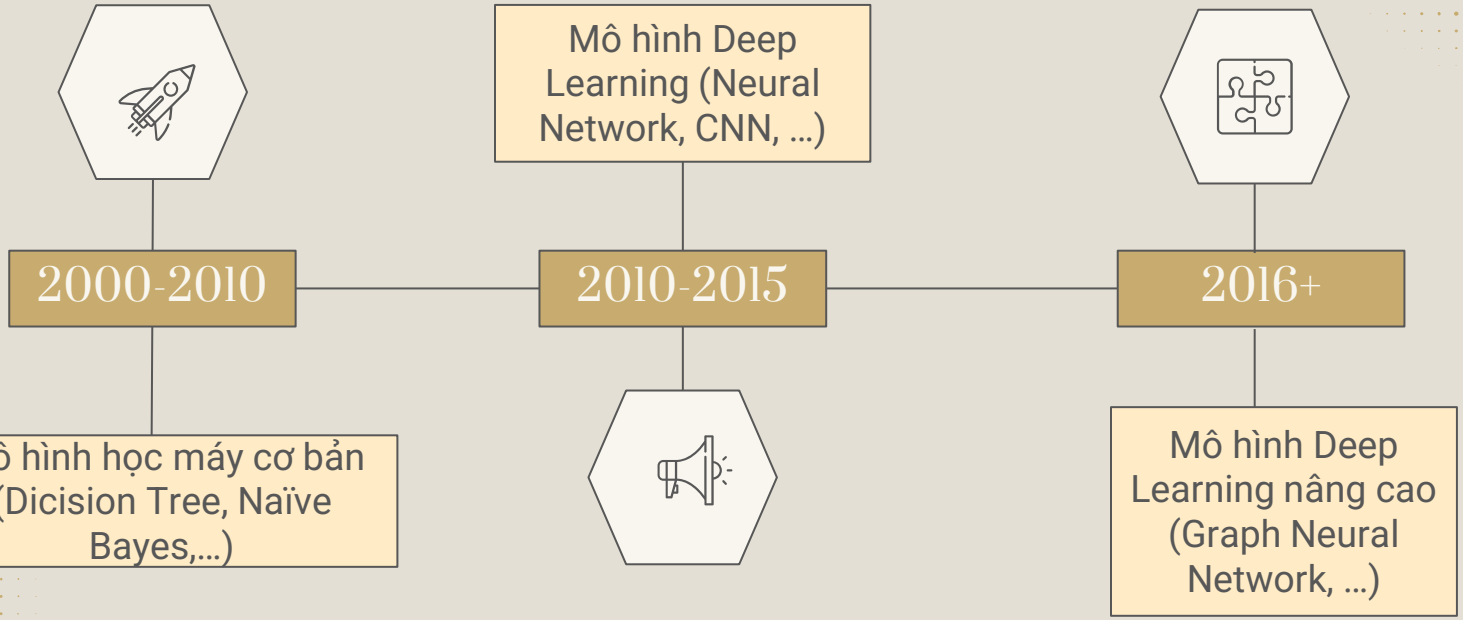
Giới thiệu bài toán

Dự đoán bệnh dựa trên triệu chứng

Bài toán dự đoán bệnh

- Các bệnh viện đã và đang phát triển các hệ thống quản lý hồ sơ bệnh nhân (lịch sử bệnh, triệu chứng, nhân khẩu học,...).
- Nhu cầu xây dựng các mô hình phân loại giúp dự đoán bệnh dựa trên triệu chứng ngày càng tăng.

Khảo sát các mô hình



Lu, H., & Uddin, S. (2023, April). Disease Prediction Using Graph Machine Learning Based on Electronic Health Data: A Review of Approaches and Trends. In *Healthcare* (Vol. 11, No. 7, p. 1031). MDPI.

Ưu điểm
Graph-based Model

- Học được mối quan hệ giữa các bệnh nhân, giữa các bệnh có cùng triệu chứng
  =>Hỗ trợ nghiên cứu thông tin của từng nhóm bệnh.

- Embedding được thông tin của tập bệnh nhân cũ và triệu chứng
  => Dễ dàng dự đoán bệnh cho bệnh nhân mới.

Ghorbani, M., Kazi, A., Baghshah, M. S., Rabiee, H. R., & Navab, N. (2022). RA-GCN: Graph convolutional network for disease prediction problems with imbalanced data. *Medical Image Analysis*, *75*, 102272.

Đề tài bao gồm

Nội dung:

- Xây dựng mô hình dự đoán bệnh dựa trên triệu chứng.

Phạm vi:

- Nhóm model Graph-based.

Mục tiêu:

- Xây dựng nhiều loại đồ thị với các node khác nhau (bệnh nhân, bệnh, triệu chứng,...).
- Xây dựng nhiều loại mô hình Graph-based:
- Graph Neural Network, GraphSage, Graph Attention,...
- Xây dựng các mô hình dự đoán với độ chỉnh xác cao về các metric phân loại: F1 score, Recall, Precision,...

02

Dữ liệu sử dụng

Khai phá dữ liệu EMR của Ấn Độ

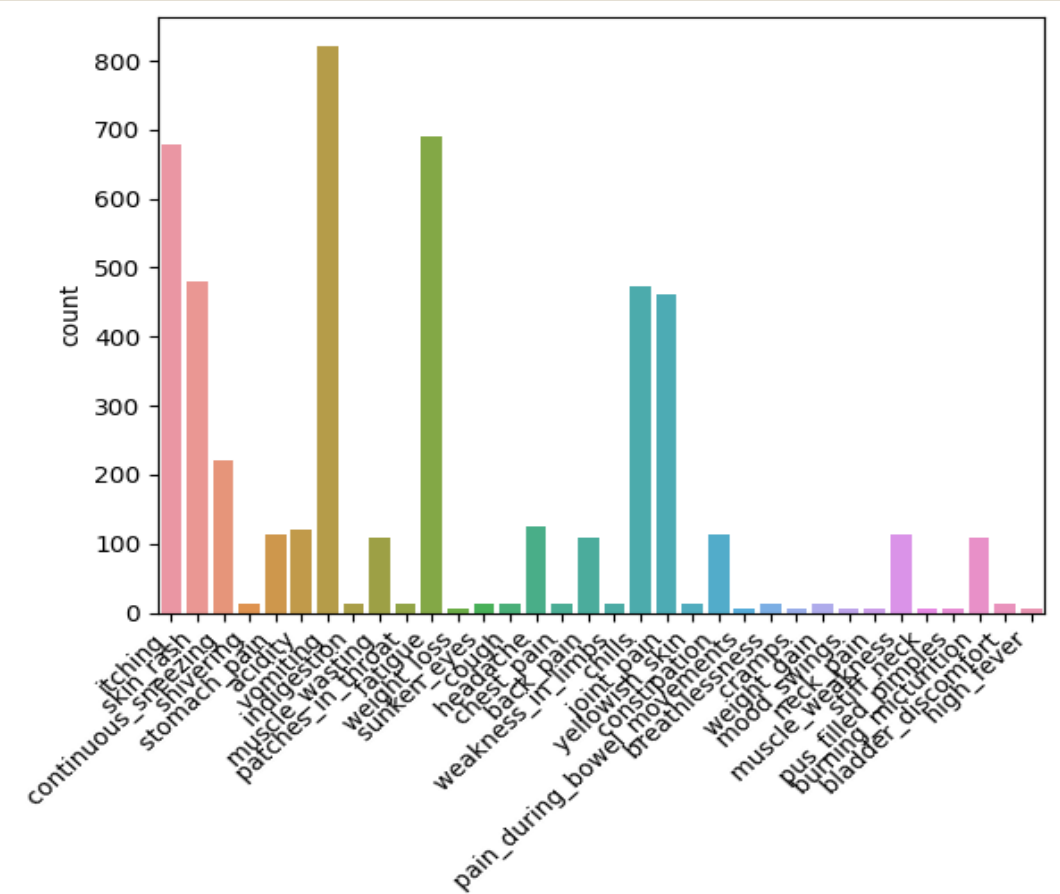
Disease Symptom Prediction

- 4920 bản ghi.
- 18 trường dữ liệu.
- 41 loại bệnh khác nhau, mỗi bệnh 120 bản ghi

	Disease	Symptom_1	Symptom_2	Symptom_3	Symptom_4
0	Fungal infection	itching	skin_rash	nodal_skin_eruptions	dischromic _patches
1	Fungal infection	skin_rash	nodal_skin_eruptions	dischromic _patches	NaN
2	Fungal infection	itching	nodal_skin_eruptions	dischromic _patches	NaN
3	Fungal infection	itching	skin_rash	dischromic _patches	NaN
4	Fungal infection	itching	skin_rash	nodal_skin_eruptions	NaN
5	Fungal infection	skin_rash	nodal_skin_eruptions	dischromic _patches	NaN
6	Fungal infection	itching	nodal_skin_eruptions	dischromic _patches	NaN

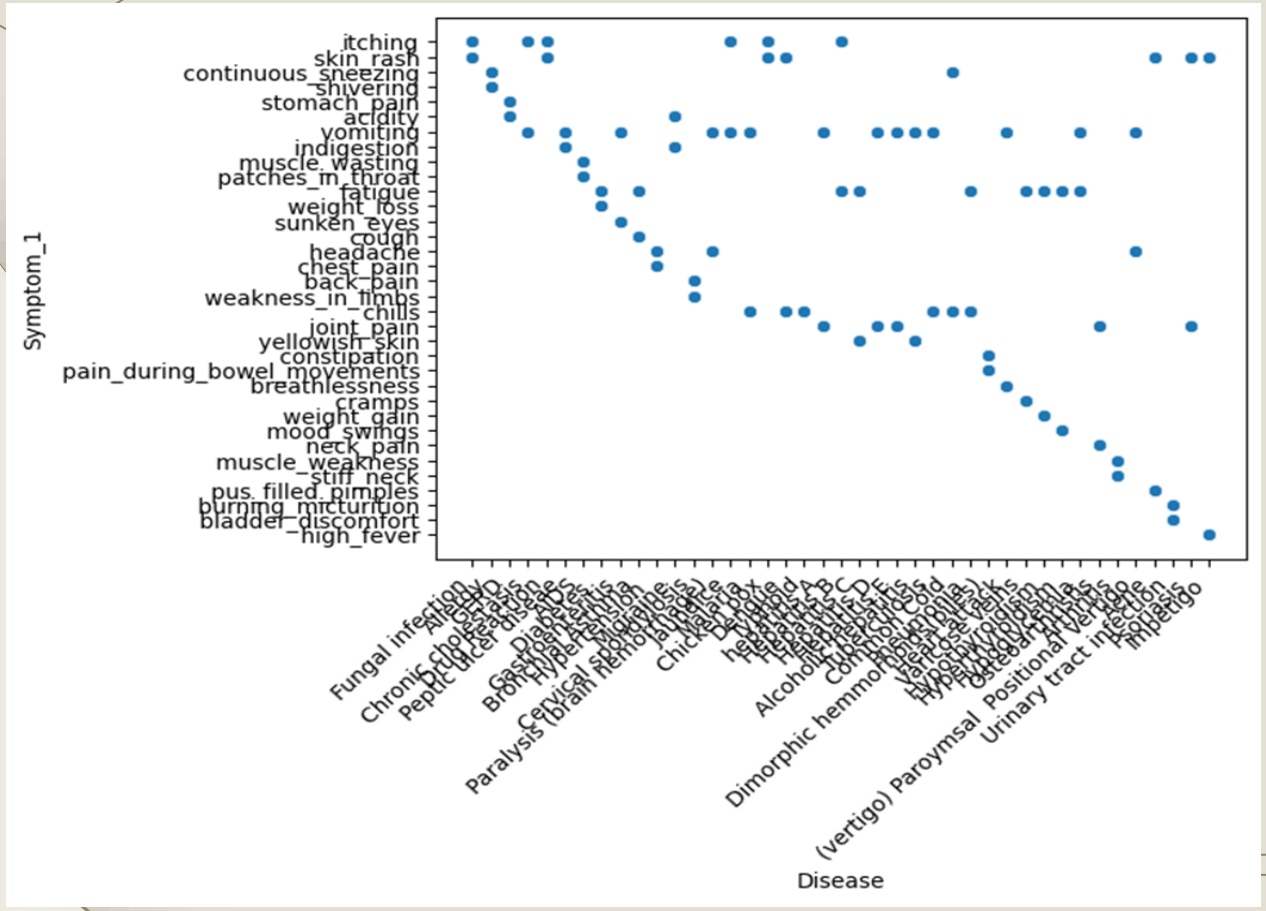
```
Fungal infection                                 120
Hepatitis C                                      120
Hepatitis E                                      120
Alcoholic hepatitis                              120
Tuberculosis                                     120
Common Cold                                      120
Pneumonia                                        120
Dimorphic hemmorhoids(piles)                     120
Heart attack                                     120
Varicose veins                                   120
Hypothyroidism                                   120
Hyperthyroidism                                  120
Hypoglycemia                                     120
Osteoarthristis                                  120
Arthritis                                        120
(vertigo) Paroymsal  Positional Vertigo          120
Acne                                             120
Urinary tract infection                          120
Psoriasis                                        120
Hepatitis D                                      120
Hepatitis B                                      120
Allergy                                          120
```

EDA (1)



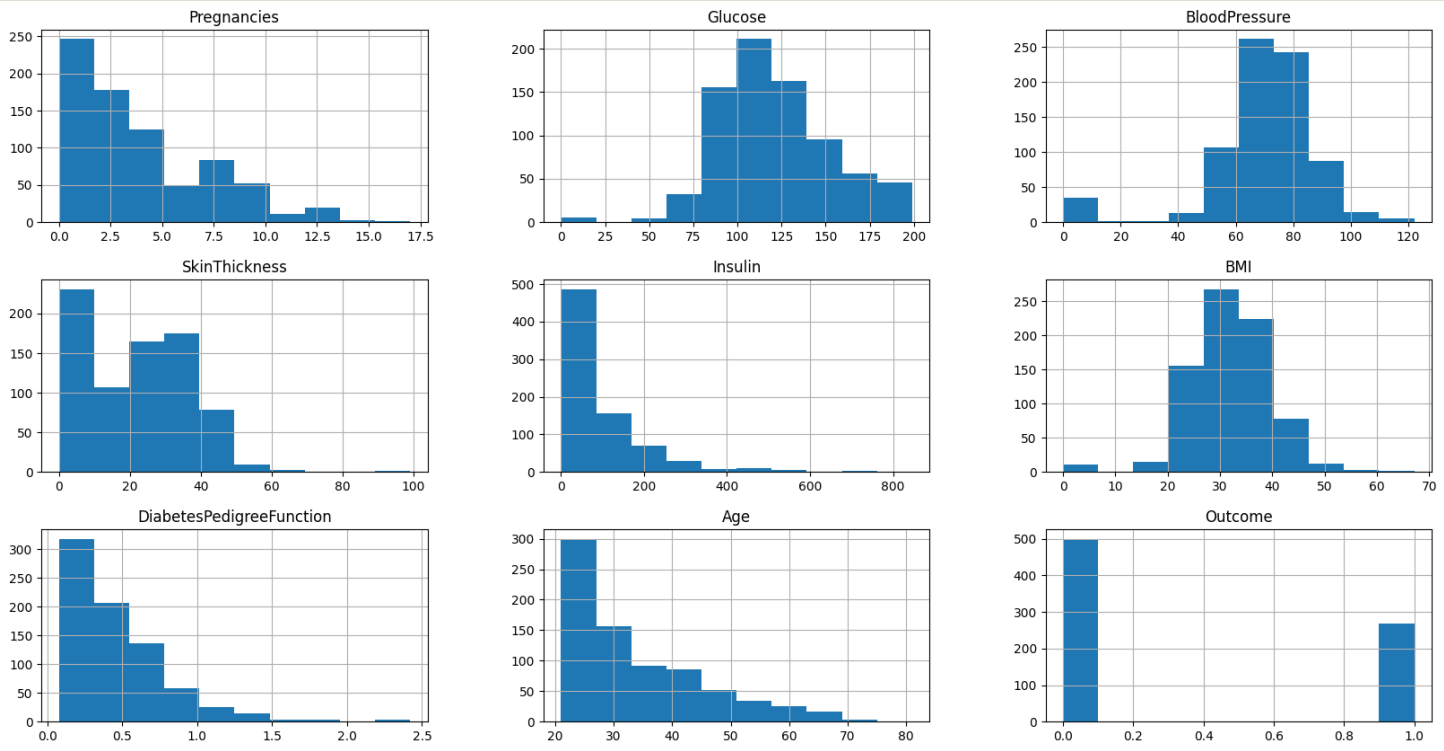
- 133 loại triệu chứng khác nhau.
- Phân bố các triệu chứng không đều.

EDA (2)



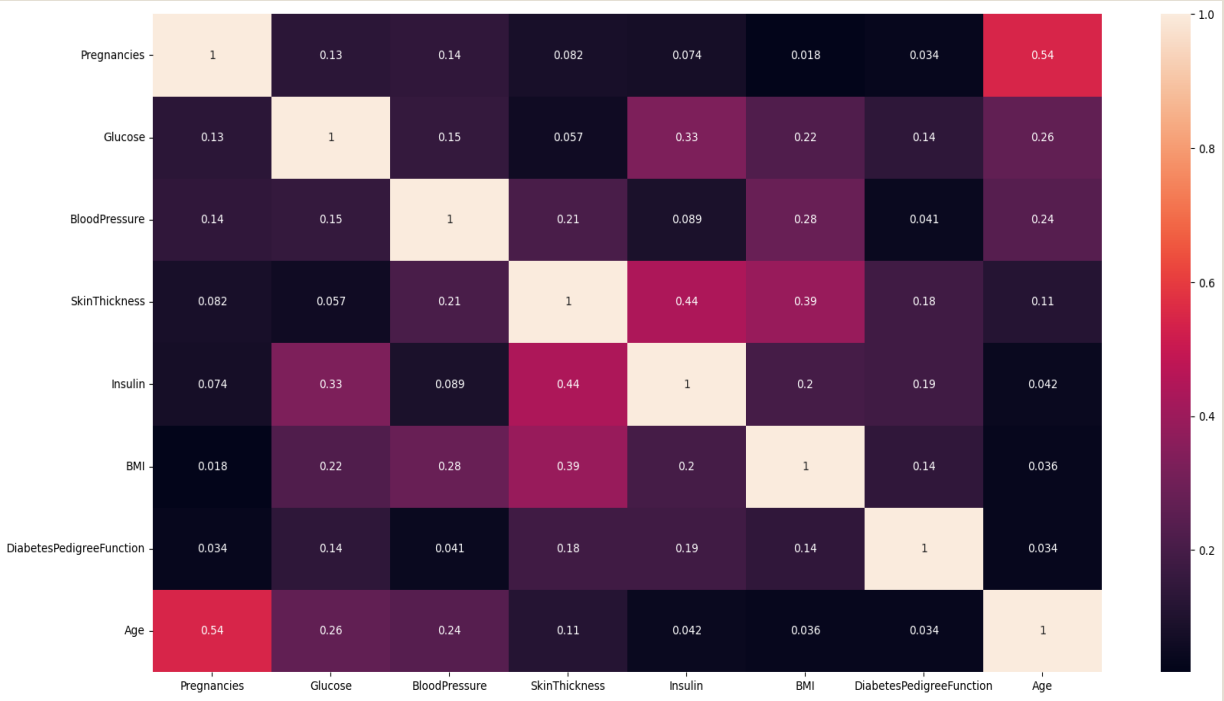
- Mỗi bệnh thường có 2 triệu chứng nổi bật
- Mội vài triệu chứng phổ biến ở các bệnh (vomiting, skin rash,..)

Pima Indian Diabetes (1)



- 768 bản ghi, 9 trường dữ liệu.
- Tỷ lệ nhãn 5:2 .

Pima Indian Diabetes (2)



Tuổi có tương quan khá cao với các biến khác trong tập dữ liệu.

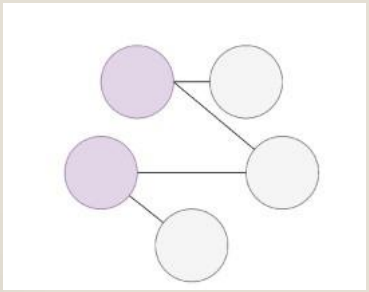
=> Tiến hành binning tuổi thành các khoảng và hình thành liên kết đồ thị giữa các nhóm tuổi.

03

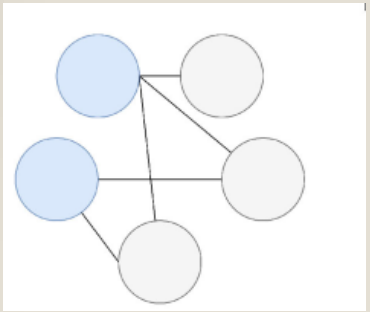
Mô hình nghiên cứu

Xây dựng các loại đồ thị khác nhau
Xây dựng Node Representation dựa trên TF-IDF
Xây dựng Mô hình Graph-based đề xuất

Các cách xây dựng đồ thị



Loại 1: Đồ thị người bệnh



Loại 2: Đồ thị bệnh

Các cách xây dựng kết nối giữa 2 người bệnh

Hướng 1: Tạo cạnh giữa 2 người bệnh nếu họ mắc cùng một loại bệnh (Để tránh data Leakage thì chỉ sử dụng tập Train để xây đồ thị).
Hướng 2: Tạo cạnh giữa 2 người bệnh nếu họ có 2 triệu chứng giống nhau. (Có thể kết hợp với đồ thị theo hướng 1 để tạo đồ thị mới).
Hướng 3: Tạo cạnh giữa 2 người bệnh nếu họ có độ tương đồng đo theo cosine similarity trên ngưỡng γ.

Các cách xây dựng kết nối giữa 2 bệnh

Hướng 1: Tạo kết nối giữa 2 bệnh nếu chúng có chung 2 triệu chứng nổi bật nhất.

Xây dựng Node Feature (l)

	Disease	Symptom_1	Symptom_2	Symptom_3
0	0	5	3	5
1	1	7	3	6
2	2	3	2	2
3	3	5	3	4
4	4	4	5	3
5	5	2	4	5
6	6	7	4	4
7	7	5	3	7
8	8	1	3	4
9	9	1	5	3
10	10	4	3	4

	Symptom	weight
0	itching	1
1	skin_rash	3
2	nodal_skin_eruptions	4
3	continuous_sneezing	4
4	shivering	5
5	chills	3
6	joint_pain	3
7	stomach_pain	5
8	acidity	3
9	ulcers_on_tongue	4

Sử dụng trọng số cho từng triệu chứng để xây dựng Node Feature cho đồ thị người bệnh

Xây dựng Node Feature (2)

Đề xuất xây dựng Feature cho đồ thị bệnh dựa trên công thức TF-IDF.

1) Thống kê 3 triệu chứng phổ biến nhất cho từng bệnh.
2) TF: Thống kê số người mắc từng loại bệnh sẽ bị triệu chứng phổ biến tương ứng.
3) IDF: Thống kê trên toàn tập số người bị các triệu chứng phổ biến.
4) Tính trọng số bằng cách lấy TF * 1/IDF.

	Disease	Symptom_1	Symptom_2	Symptom_3
0	0	vomiting	headache	nausea
1	1	high_fever	muscle_wasting	patches_in_throat
2	2	skin_rash	pus_filled_pimples	blackheads
3	3	vomiting	yellowish_skin	abdominal_pain
4	4	continuous_sneezing	shivering	chills
5	5	muscle_weakness	stiff_neck	swelling_joints
6	6	high_fever	fatigue	cough
7	7	neck_pain	back_pain	weakness_in_limbs
8	8	itching	skin_rash	fatigue
9	9	itching	vomiting	yellowish_skin

	Disease	Symptom_1	Symptom_2	Symptom_3
0	0	0.059561	0.100529	0.099476
1	1	0.0837	0.05	0.05
2	2	0.145038	0.05	0.05
3	3	0.059561	0.125	0.110465
4	4	0.486486	0.05	0.135338
5	5	0.487179	0.5	0.5
6	6	0.0837	0.055901	0.191489
7	7	0.5	0.473684	0.05
8	8	0.168142	0.145038	0.059006
9	9	0.168142	0.059561	0.125
10	10	0.513514	0.142857	0.059006
11	11	0.145038	0.142857	0.166667

Các bước học của mô hình

1. Prepare:

Input node representations được đi qua 1 lớp Multi Layer Perceptron để biểu diễn thông tin message.

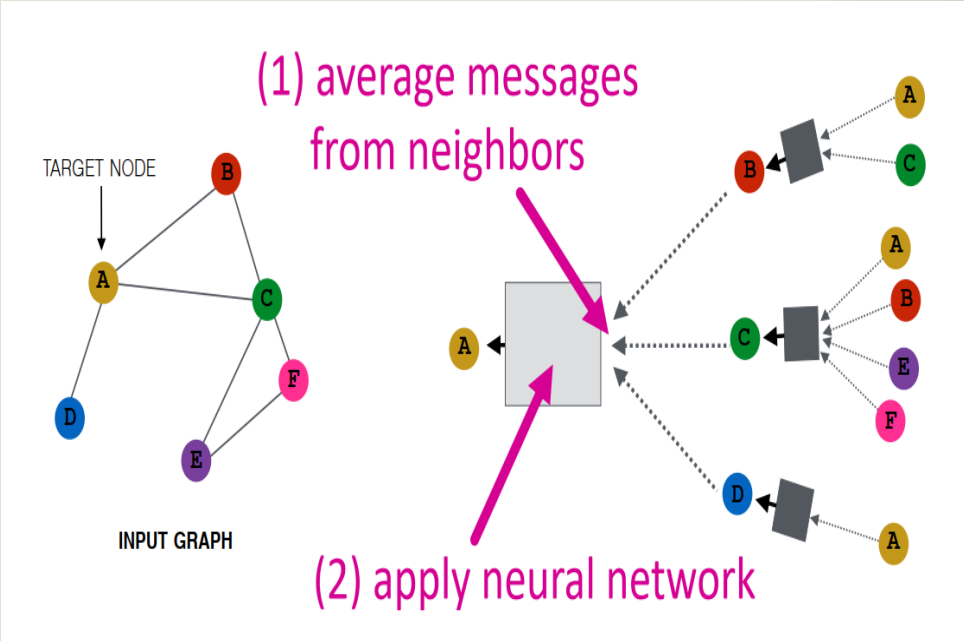
2. Aggregate:

Thông tin của mỗi node lân cận (bậc 1) sẽ được kết hợp lại để giúp mô hình học được liên kết giữa các node .

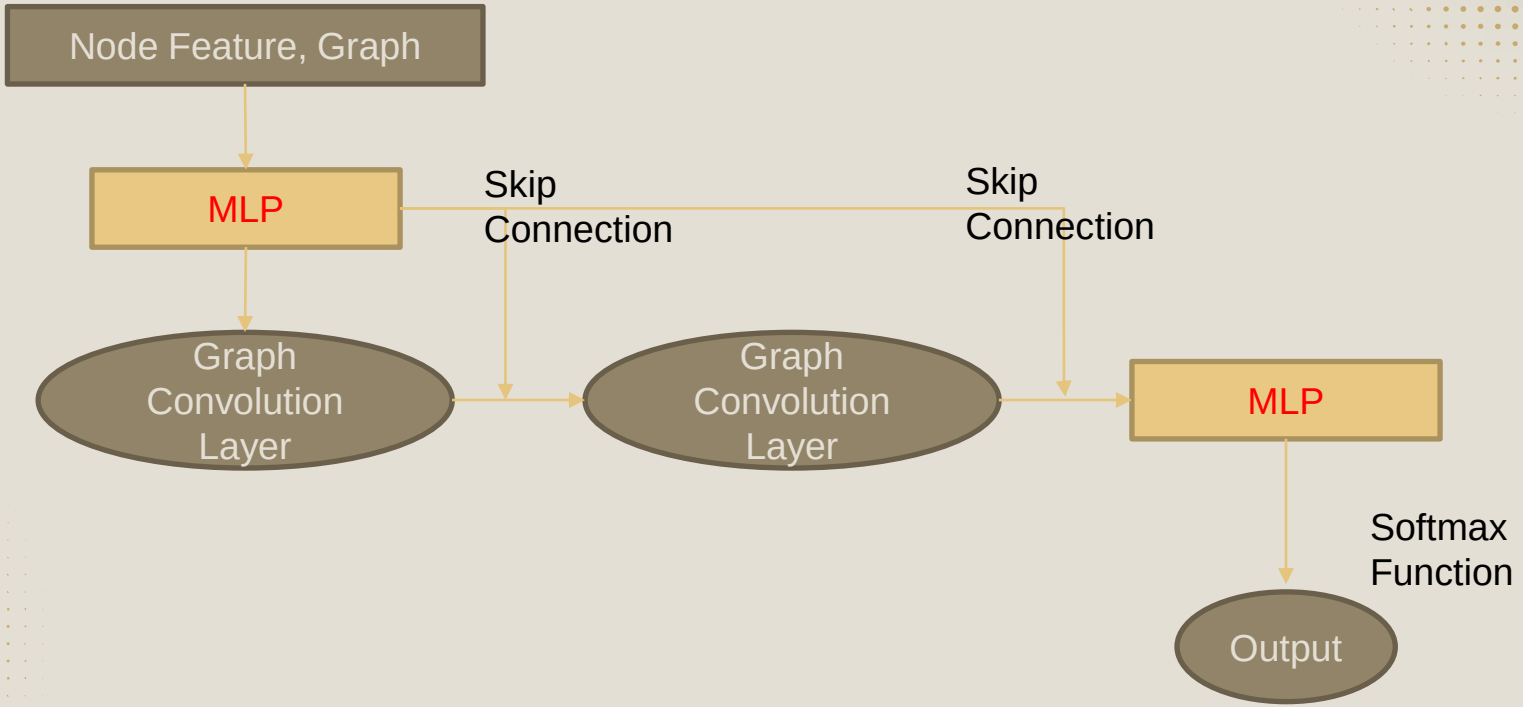
Một số cách để kết hợp thông tin: Sum, Max, Mean, Attention, Multi Head Attention.

3. Update:

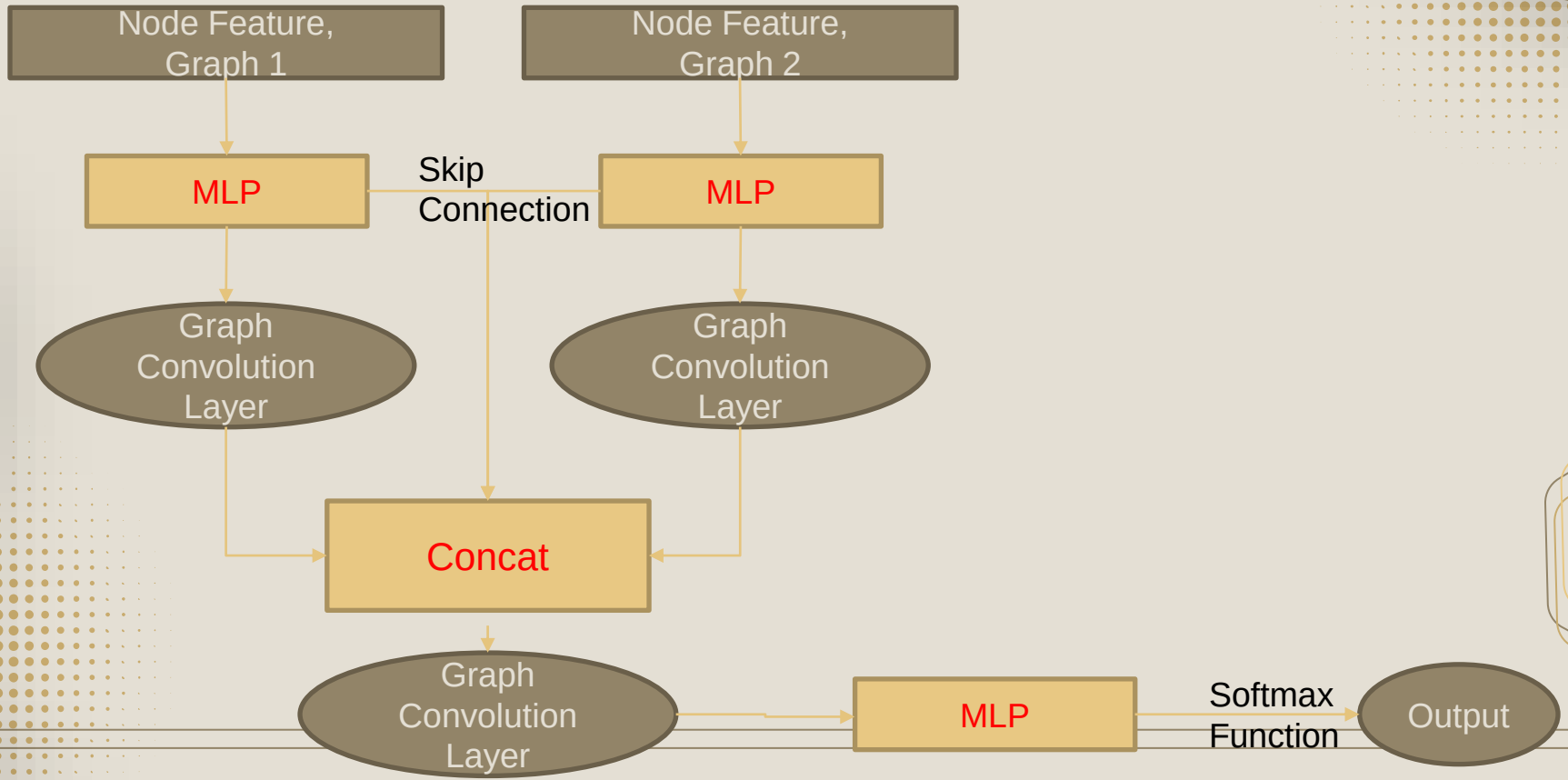
Node representations sẽ được cập nhật bởi thông tin kết hợp của các lân cận. Cập nhật có thể được dung bởi lớp GRU hoặc MLP.



Mô hình đề xuất (Đơn đồ thị Input)



Mô hình đề xuất (Đa đồ thị Input)



Cách thức huấn luyện

Giả thuyết

Việc giảm kích thước đồ thị Input bằng cách sample một lượng nhỏ các kết nối (5%, 10%, 20%, 30%…) không làm giảm hiệu suất mô hình nếu kết hợp thông tin bằng phương thức Attention.

Giải thích

- Dựa vào dữ liệu, mỗi người chỉ cùng nhóm bệnh với 120 người khác (1/41) dữ liệu. Đối với từng nhóm bệnh chỉ cần một nhóm nhỏ số người đại diện.

- Cơ chế Attention giúp mô hình học được các liên kết quan trọng trong đồ thị. Một node sẽ học được đâu là liên kết quan trọng với các node lân cận. Do đó việc bỏ có xác suất các cạnh không ảnh hưởng tới hiệu suất.

=> Kết quả thực nghiệm ở phần dưới chứng minh độ hiệu quả của cách thức huấn luyện

04

Kết quả chạy mô hình

So sánh, đánh giá kết quả chạy các mô hình đề xuất.
So sánh với các kết quả đã được công bố.

Disease Symptom Prediction

Chia tập dữ liệu Train – Valid – Test: 70/10/20

Model	Accuracy	Macro F1	Macro Recall	Macro Precision
BaseLine Random Forest	0.91±0.01	0.90±0.01	0.90±0.01	0.95±0.01
Hướng 1 (Mean + GRU)	0.99±0.01	0.99±0.01	0.99±0.01	0.99±0.01
Hướng 2 (Sum + Concat)	0.98±0.01	0.98±0.01	0.99±0.01	0.98±0.01
Hướng 3 (Attention + GRU + sample 20%)	0.95±0.01	0.94±0.01	0.95±0.01	0.95±0.01
Hướng 3 (MultiHeadAttention + concat + sample 5%)	0.99±0.01	0.99±0.01	1±0.01	0.99±0.01
Đa đồ thị (Mean + GRU)	1±0.01	1±0.01	1±0.01	1±0.01

Nhận xét: Việc giảm kích thước đồ thị và dung cơ chế Attention cho kết quả tốt như giả thuyết.

Pima Indian Diabetes

Chia tập dữ liệu Train – Valid – Test: 70/10/20

- Tự xây dựng triệu chứng bằng cách Binning trường Age => Age Range.
- Đánh Class weight cho 2 nhãn để xử lý việc mất cân bằng dữ liệu (tỷ lệ 5:2).

Model	Accuracy	Macro F1	Macro Recall	Macro Precision
BaseLine KNN	0.73±0.01	0.73±0.01	0.73±0.01	0.74±0.01
Hướng 1 (Mean + GRU)	0.77±0.01	0.76±0.01	0.76±0.01	0.79±0.01
Hướng 2 (Sum + Concat)	0.71±0.01	0.65±0.01	0.64±0.01	0.79±0.01
Hướng 3 (Attention + GRU)	0.77±0.01	0.77±0.01	0.77±0.01	0.79±0.01

Nhận xét: Việc binning dữ liệu liên tục về rời rạc và coi như triệu chứng để tiến hành xây dựng đồ thị chưa quá hiệu quả.

05

Kết luận

Kết quả đạt được và hạn chế của đề tài.
Hướng phát triển đề tài .

Kết quả - Hạn chế

Kết quả:

- Hiểu và xây dựng được các mô hình Graph-based như GraphSage, Graph Attention,... với một hoặc nhiều đồ thị làm input.
- Đề xuất các cách xây dựng đồ thị khác nhau.(Đồ thị bệnh, đồ thị triệu chứng, xây dựng bộ feature cho bệnh dựa trên áp dụng độ đo TF-IDF với các triệu chứng nổi bật,...)
- Đề xuất mô hình Graph-based MultiHeadAttention với cơ chế sample các cạnh giúp giảm kích thước đồ thị mà vẫn đảm bảo hiệu suất phân loại.

Hạn chế:

- Chưa xây dựng cách thiết lập trọng số cạnh cho các đồ thị.

Hướng phát triển đề tài

- Bổ sung thêm dữ liệu (Crawl Data) để đánh giá hiệu quả của mô hình với bộ dữ liệu lớn.
- Xây dựng nhiều loại đồ thị với các node khác nhau (Đồ thị gồm các node bệnh nhân, bệnh, triệu chứng liên kết với nhau).

Thank you