ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc Lập - Tự Do - Hạnh Phúc

THÔNG TIN ĐĂNG KÝ ĐỀ TÀI LUẬN VĂN THẠC SỸ

1. **Tên đề tài (ghi IN HOA):**
   - Tên tiếng Việt: PHÂN TÍCH THÀNH PHẦN CHÍNH XÁC SUẤT TỔNG QUÁT CHO DỮ LIỆU TƯƠNG QUAN
   - Tên tiếng Anh: GENERALIZED PROBABILISTIC PRINCIPAL COMPONENT ANALYSIS OF CORRELATED DATA
   - Hướng đề tài luận văn:
     - Hướng nghiên cứu □
     - Định hướng nghiên cứu □
     - Định hướng ứng dụng □
   - Số tín chỉ:
2. **Ngành học và Mã ngành:**
   - Khoa học máy tính: 8480101 □
   - Công nghệ Thông tin: 8480201 □
3. **Cán bộ hướng dẫn:** *(định dạng 2 cột nếu có 2 CBHD)*
   - Họ tên:
   - Email:
   - Điện thoại:
   - Đơn vị công tác:
4. **Thời gian thực hiện: 6 tháng.** Từ tháng …../20…..
5. **Học viên thực hiện:**

- Họ tên:
- Mã số: Khóa: Đợt:
- Email: Điện thoại:

Xác nhận của CBHD
(Ký tên và ghi rõ họ tên)

TP. HCM, ngày….tháng …..năm 20….
Học viên
(Ký tên và ghi rõ họ tên)

ĐỀ CƯƠNG ĐỀ TÀI LUẬN VĂN THẠC SỸ

I. Nội dung

1.1. Giới thiệu đề tài

1.1.1. Đặt vấn đề

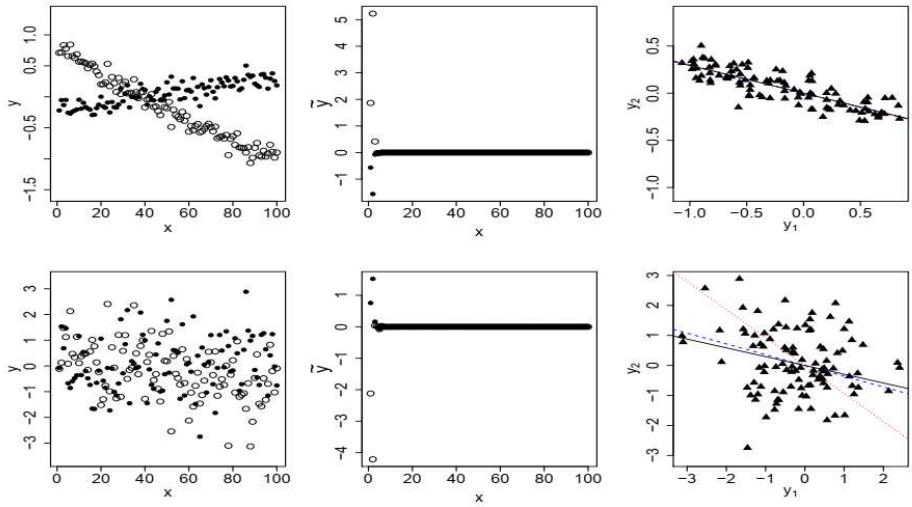
Phân tích thành phần chính (PCA) là một trong những cách tiếp cận lâu đời nhất và được biết đến rộng rãi nhất để giảm số chiều dữ liệu. Nó đã được sử dụng trong nhiều ứng dụng, bao gồm phân tích dữ liệu, hồi quy, phân tích chuỗi thời gian, xử lý hình ảnh,…

1.1.2. Tình hình nghiên cứu

- Giải pháp phổ biến nhất của PCA là tìm phép chiếu tuyến tính biến tập hợp các biến tương quan ban đầu thành một tập hợp các biến mới không tương quan bởi phương sai cực đại của các biến mới sinh ra (Jolliffe, 2011). Giải pháp này, mặc dù được sử dụng rộng rãi trong thực tế, nhưng nó thiếu mô tả xác suất của dữ liệu.[2]
- Năm 1999, Tipping và Bishop lần đầu tiên đã giới thiệu giải pháp xác suất cho PCA, trong đó nhóm tác giả xem mỗi biến thành phần là một mô hình Gaussian và đạt được PCA (các trục chính) bởi nghiệm của bài toán ước lượng hợp lý cực đại nhưng trong đó các biến tương quan bị bỏ qua. Cách tiếp cận này được gọi là phân tích thành phần chính xác suất (PPCA). [3]
- Năm 2020, Mengyang Gu và Weining Shen đã đưa ra giải pháp phân tích thành phần chính xác suất tổng quát (GPPCA) như một mở rộng của PPCA cho nhiều thành phần dữ liệu đầu ra tương quan, trong đó mỗi thành phần là một quá trình Gaussian.[1]

1.1.3. Input/Output

- Input: Cho n điểm dữ liệu trong không gian p chiều, p thành phần được biểu diễn dạng véctơ $x = (x_1, x_2, \dots, x_p)^T$.
- Output: Số chiều dữ liệu được rút gọn xuống thành k thành phần, được biểu diễn dạng véc tơ $y(x) = (y_1(x), y_2(x), \dots, y_k(x))^T$ với $k < p$.



1.1.4. Tại sao chọn đề tài

Phân tích thành phần chính (PCA) là một trong những kỹ thuật quan trọng trong học máy và xử lí dữ liệu nhằm giảm số chiều dữ liệu.

1.1.5. Tính khả thi của đề tài

Mặc dù không thời sự nhưng lợi thế của GPPCA xét về mức độ phù hợp thực tế, ước tính độ chính xác và tính thuận tiện trong tính toán. Thí nghiệm trên bộ dữ liệu thực và dữ liệu mô phỏng đạt đạt hiệu quả cao hơn so với các cách tiếp cận khác.

1.2. Mục tiêu của đề tài

- Đưa ra công thức tính toán phân phối biên của các thành phần bằng phương pháp ước lượng hợp lý cực đại của ma trận hiệp phương sai.
- Đưa ra ước lượng tham số và dự đoán phân phối.
- Đưa ra cấu trúc của kỳ vọng khi thêm vào các biến tương quan.

1.3. Đối tượng nghiên cứu

1.4. Phương pháp nghiên cứu

- Phương pháp thu thập số liệu: dựa trên bộ dữ liệu thực RMSE
- Phương pháp mô phỏng dữ liệu mô phỏng
- Phương pháp phân tích so sánh trên 2 bộ dữ liệu để so sánh tính hiệu quả của phương pháp GPPCA là cao hơn so với các cách tiếp cận khác trước đây.

1.5. Phạm vi nghiên cứu

Thời gian nghiên cứu 6 tháng, dựa trên bộ dữ liệu RMSE và so sánh với các phương pháp trước đây.

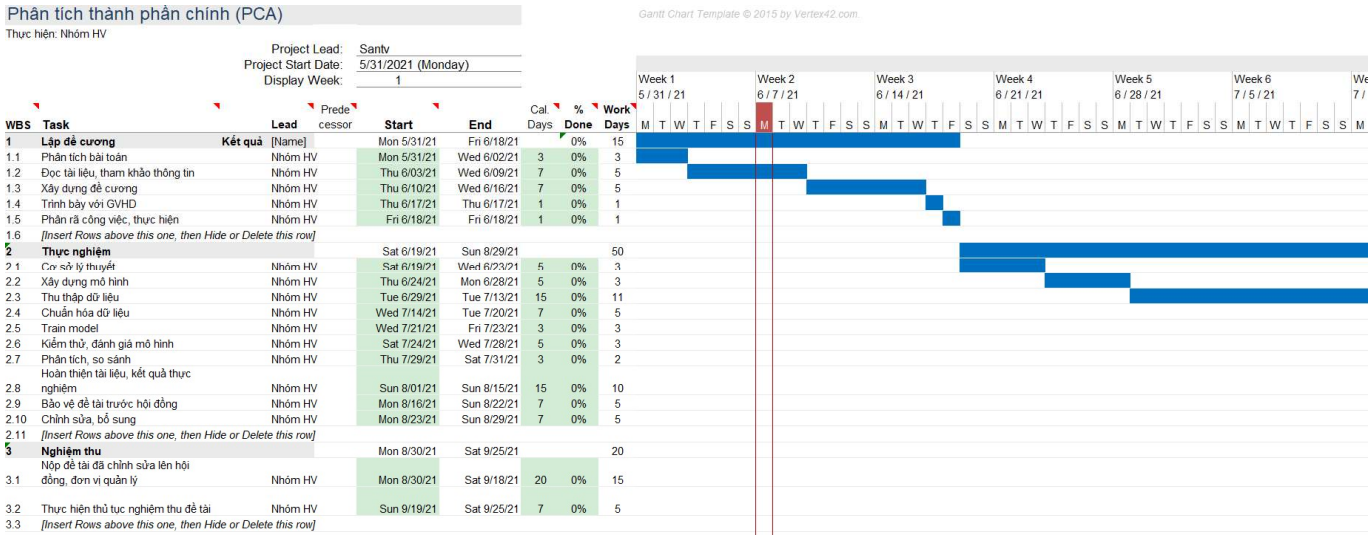
1.6. Nội dung nghiên cứu

- Xác suất tổng quát phân tích thành phần chính (GPPCA) cho mô hình dữ liệu cho nhiều thành phần tương quan, trong đó mỗi thành phần được mô hình hóa bằng một quá trình Gaussian. Bằng phương pháp ước lượng hợp lý cực đại đưa ra công thức tính toán phân phối biên của các thành phần.
- Dựa vào ma trận biểu thị độ chính xác trong ước lượng hợp lý cực đại đưa ra ước lượng tham số và dự đoán phân phối.
- Dựa vào phương pháp ước lượng hợp lý cực đại cho kỳ vọng cho các biến tương quan.

1.7. Kết quả, sản phẩm dự kiến

- Đưa ra công thức tổng quát hơn để xác định các tham số.
- Chứng minh hiệu quả của GPPCA xét về mức độ phù hợp thực tế, ước tính độ chính xác và tính thuận tiện trong tính toán.

II. Kế hoạch



WBS	Task	Lead	Predecessor	Start	End	Cal. Days	% Done	Work Days
1	**Lập đề cương**	[Name]		Mon 5/31/21	Fri 6/18/21		0%	15
1.1	Phân tích bài toán	Nhóm HV		Mon 5/31/21	Wed 6/02/21	3	0%	3
1.2	Đọc tài liệu, tham khảo thông tin	Nhóm HV		Thu 6/03/21	Wed 6/09/21	7	0%	5
1.3	Xây dựng đề cương	Nhóm HV		Thu 6/10/21	Wed 6/16/21	7	0%	5
1.4	Trình bày với GVHD	Nhóm HV		Thu 6/17/21	Thu 6/17/21	1	0%	1
1.5	Phân rã công việc, thực hiện	Nhóm HV		Fri 6/18/21	Fri 6/18/21	1	0%	1
1.6	*[Insert Rows above this one, then Hide or Delete this row]*							
2	**Thực nghiệm**			Sat 6/19/21	Sun 8/29/21			50
2.1	Cơ sở lý thuyết	Nhóm HV		Sat 6/19/21	Wed 6/23/21	5	0%	3
2.2	Xây dựng mô hình	Nhóm HV		Thu 6/24/21	Mon 6/28/21	5	0%	3
2.3	Thu thập dữ liệu	Nhóm HV		Tue 6/29/21	Tue 7/13/21	15	0%	11
2.4	Chuẩn hóa dữ liệu	Nhóm HV		Wed 7/14/21	Tue 7/20/21	7	0%	5
2.5	Train model	Nhóm HV		Wed 7/21/21	Fri 7/23/21	3	0%	3
2.6	Kiểm thử, đánh giá mô hình	Nhóm HV		Sat 7/24/21	Wed 7/28/21	5	0%	3
2.7	Phân tích, so sánh	Nhóm HV		Thu 7/29/21	Sat 7/31/21	3	0%	2
2.8	Hoàn thiện tài liệu, kết quả thực nghiệm	Nhóm HV		Sun 8/01/21	Sun 8/15/21	15	0%	10
2.9	Bảo vệ đề tài trước hội đồng	Nhóm HV		Mon 8/16/21	Sun 8/22/21	7	0%	5
2.10	Chỉnh sửa, bổ sung	Nhóm HV		Mon 8/23/21	Sun 8/29/21	7	0%	5
2.11	*[Insert Rows above this one, then Hide or Delete this row]*							
3	**Nghiệm thu**			Mon 8/30/21	Sat 9/25/21			20
3.1	Nộp đề tài đã chỉnh sửa lên hội đồng, đơn vị quản lý	Nhóm HV		Mon 8/30/21	Sat 9/18/21	20	0%	15
3.2	Thực hiện thủ tục nghiệm thu đề tài	Nhóm HV		Sun 9/19/21	Sat 9/25/21	7	0%	5
3.3	*[Insert Rows above this one, then Hide or Delete this row]*							

Tài liệu tham khảo

[1]- Mengyang Gu and Weining Shen. Generalized probabilistic principal component analysis of correlated data. *Journal of Machine Learning Research* 21:1- 41, 2020.

[2]- Ian Jolliffe. *Principal component analysis*. Springer, 2011.

[3]- Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611- 622, 1999.