ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

VŨ THỊ NGUYỆT

NGHIÊN CỨU PHƯƠNG PHÁP GIẢM CHIỀU DỮ LIỆU VỚI PCA VÀ MỘT SỐ ỨNG DỤNG

Chuyên ngành: Khoa học máy tính
Mã số: 8 48 01 01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Giáo viên hướng dẫn: TS. Đàm Thanh Phương

THÁI NGUYÊN - 2021

LỜI CAM ĐOAN

Tôi xin cam đoan: Luận văn thạc sỹ chuyên ngành Khoa học máy tính, tên đề tài "Nghiên cứu phương pháp giảm chiều dữ liệu với PCA và một số ứng dụng" là công trình nghiên cứu, tìm hiểu và trình bày do tôi thực hiện dưới sự hướng dẫn khoa học của TS. Đàm Thanh Phương, Trường Đại học Công nghệ Thông tin và Truyền thông - Đại học Thái Nguyên.

Kết quả tìm hiểu, nghiên cứu trong luận văn là hoàn toàn trung thực, không vi phạm bất cứ điều gì trong luật sở hữu trí tuệ và pháp luật Việt Nam. Nếu sai, tôi hoàn toàn chịu trách nhiệm trước pháp luật.

Tất cả các tài liệu, bài báo, khóa luận, công cụ phần mềm của các tác giả khác được sử dụng lại trong luận văn này đều được chỉ dẫn tường minh về tác giả và đều có trong danh mục tài liệu tham khảo.

Thái Nguyên, ngày 18 tháng 2 năm 2021.

Tác giả luận văn

Vũ Thị Nguyệt

LỜI CẢM ƠN

Tác giả xin chân thành cảm ơn TS Đàm Thanh Phương, trường Đại học Công nghệ thông tin và truyền thông - Đại học Thái Nguyên, là giáo viên hướng dẫn khoa học đã hướng dẫn tác giả hoàn thành luận văn này, xin được cảm ơn các thầy, cô giáo trường Đại học công nghệ thông tin và truyền thông nơi tác giả theo học và hoàn thành chương trình cao học đã nhiệt tình giảng dạy và giúp đỡ.

Xin cảm ơn trường THPT Cẩm Phả - Quảng Ninh nơi tác giả công tác đã tạo mọi điều kiện thuận lợi để tác giả có thời gian, tâm trí để hoàn thành nhiệm vụ nghiên cứu và chương trình học tập.

Và cuối cùng xin cảm ơn gia đình, bạn bè, đồng nghiệp đã động viên, giúp đỡ tác giả trong suốt thời gian học tập, nghiên cứu và hoàn thành luận văn này.

Xin chân thành cảm ơn.

Thái Nguyên, ngày 18 tháng 2 năm 2021

Tác giả luận văn

Vũ Thị Nguyệt

DANH SÁCH HÌNH VẼ

1.1 Kiến trúc của autoendcoder và hàm loss của nó 12

2.1 Ví dụ về các điểm dữ liệu 2 chiều có phương sai theo chiều thứ nhất lớn. 20
2.2 Ví dụ về các điểm dữ liệu 2 chiều có phương sai theo cả 2 chiều đều lớn. 22
2.3 Mô tả ý tưởng đổi hệ tọa độ của PCA. Dữ liệu được biểu diễn qua hệ cơ sở mới thỏa mãn mong muốn về sự chênh lệch phương sai giữa các thành phần. 22
2.4 Minh họa dữ liệu trong hệ cơ sở trực chuẩn tìm được bằng PCA. 26
2.5 Các bước thực hiện PCA. 28

3.1 Ví dụ về ảnh của một người trong Yale Face Database. 38
3.2 Các eigenfaces tìm được bằng PCA. 40
3.3 Hàng trên: các ảnh gốc. Hàng dưới: các ảnh được tái tạo dùng khuôn mặt riêng. Ảnh ở hàng dưới có nhiễu nhưng vẫn mang những đặc điểm riêng mà mắt người có thể phân biệt được. . . . 41
3.4 PCA cho bài toán dò tìm điểm bất thường. 41
3.5 Danh mục đầu tư chính sử dụng PCA. 43
3.6 Trực quan dữ liệu sử dụng 2 thành phần chính trong PCA. . . . 45
3.7 Dữ liệu biểu diễn theo 8 chiều chính. 46
3.8 Tái tạo lại chữ số bằng 8 cơ sở PCA đầu tiên. 46
3.9 Tương quan giữa số thành phần chính giữ lại và Phương sai. . . . 47
3.10 Dữ liệu chưa có nhiễu. 48
3.11 Dữ liệu đã cộng nhiễu 49
3.12 Dữ liệu sau khi giảm chiều PCA, đã chống được nhiễu. 49

DANH MỤC KÝ HIỆU, TỪ VIẾT TẮT

$\mathbb{R}$	Tập hợp số thực.
$\mathbb{Z}$	Tập hợp số nguyên.
$\mathbb{C}$	Tập hợp số phức.
$\mathbb{R}^d$	Không gian Euclide d chiều.
C^k	Không gian các hàm có đạo hàm cấp k liên tục.
$\|\|.\|\|$	Chuẩn Euclide.
$\|\|.\|\|_F$	Chuẩn Frobenius.
PCA	Principal Component Analysis- Phân tích thành phần chính
ML	Machine Learning - Học máy.
$Trace(A)$	Vết của ma trận A.
$Span(S)$	Không gian sinh bởi hệ S.
LDA	Linear Discriminant Analysis - Phân tích biệt thức tuyến tính.
KPCA	Kernel PCA
Eigenface	Khuân mặt riêng
EigenProfolio	Danh mục đầu tư chính
MNIST	Bộ cơ sở dữ liệu chữ số viết tay.
SVD	Singular Value Decomposition - Phân tích giá trị riêng.

MỤC LỤC

Lời cam đoan ..... **i**

Lời cảm ơn ..... **ii**

Danh sách hình vẽ ..... **iii**

Danh mục ký hiệu, từ viết tắt ..... **iv**

Mở đầu **1**

Chương 1. Tổng quan học máy và bài toán giảm chiều dữ liệu ... **5**

1.1. Tổng quan về học máy **5**

1.2. Tổng quan về giảm chiều dữ liệu **10**

1.3. Nền tảng toán học **13**

Chương 2. PHƯƠNG PHÁP PCA GIẢM CHIỀU DỮ LIỆU .. **19**

2.1. Phát biểu bài toán **19**

2.2. Phân tích thành phần chính **20**

Chương 3. MỘT SỐ ỨNG DỤNG CỦA PCA **33**

3.1. Khuôn mặt riêng **33**

3.2. Dò tìm điểm bất thường **40**

3.3. Ứng dụng PCA trong tài chính **41**

3.4. Ứng dụng PCA trong trực quan hóa dữ liệu, khử nhiễu . **44**

Kết luận chung **50**

Tài liệu tham khảo **51**

Phụ lục code chương trình **53**

MỞ ĐẦU

Ngày nay, khi xã hội ngày càng phát triển, việc đưa máy tính vào sử dụng, phục vụ cho công việc đời sống của con người đã sản sinh ra một khối lượng dữ liệu lớn và phức tạp (big data), được số hóa và lưu trữ trên máy tính. Những tập dữ liệu lớn này có thể bao gồm các dữ liệu có cấu trúc, không có cấu trúc và bán cấu trúc. Đó có thể là dữ liệu thông tin bán hàng trực tuyến, lưu lượng truy cập trang web, thông tin cá nhân, thói quen hoạt động thường ngày của con người.v.v. Chúng chứa đựng nhiều thông tin quý báu mà khi khai thác hợp lý sẽ trở thành tri thức, tài sản mang lại giá trị lớn. Thách thức đặt ra cho con người là phải đưa ra các phương pháp, thuật toán và công cụ hợp lý làm sao để phân tích được lượng dữ liệu lớn như vậy.

Người ta nhận thấy máy tính có khả năng phân tích, xử lí khối dữ liệu lớn và phức tạp, tìm ra các mẫu và quy luật, vượt quá khả năng, tốc độ tính toán ghi nhớ của bộ não con người. Khái niệm học máy từ đó hình thành. Ý tưởng cơ bản của học máy là máy tính có thể học hỏi, học tự động theo kinh nghiệm [1]. Máy tính phân tích lượng lớn dữ liệu, tìm thấy các mẫu, quy tắc ẩn trong dữ liệu, sử dụng các quy tắc đó để mô tả dữ liệu mới một cách tự động và liên tục cải thiện.

Học máy có rất nhiều ứng dụng, bao gồm nhiều lĩnh vực. Máy tìm kiếm sử dụng học máy để xây dựng mối quan hệ tốt hơn giữa các cụm từ tìm kiếm và các trang web. Bằng cách phân tích nội dung của các trang web, công cụ tìm kiếm có thể xác định từ nào là cụm từ quan trọng nhất trong việc xác định một trang web nhất định và họ có thể sử dụng cụm từ này để trả thông tin kết quả phù hợp cho cụm từ tìm kiếm nhất định [2]. Công nghệ nhận dạng hình ảnh cũng sử dụng học máy để xác định các đối tượng cụ thể, chẳng hạn như khuôn mặt [5]. Đầu tiên thuật toán học máy phân tích hình ảnh có chứa một đối tượng nhất định. Nếu được cung cấp đủ hình ảnh cho quá trình này,

thuật toán có thể xác định được hình ảnh có chứa đối tượng đó hay không [3]. Ngoài ra học máy có thể được sử dụng để hiểu loại sản phẩm mà khách hàng quan tâm, bằng cách phân tích các sản phẩm trong quá khứ mà người dùng đã mua. Máy tính có thể đưa ra đề xuất các sản phẩm khách hàng có thể mua với xác suất cao [1]. Tất cả những ví dụ trên đều có nguyên tắc cơ bản giống nhau: Máy tính xử lý và học cách xác định dữ liệu, sau đó sử dụng kiến thức này để đưa ra quyết định về dữ liệu trong tương lai.

Tùy theo loại dữ liệu đầu vào, thuật toán học máy có thể được chia thành học có giám sát và học không giám sát. Trong học có giám sát, dữ liệu đầu vào đã có nhãn và đi kèm với một cấu trúc đã biết [1], [5]. Dữ liệu đầu vào được gọi là dữ liệu huấn luyện. Thuật toán thường có nhiệm vụ tạo ra một mô hình có thể dự đoán một số thuộc tính từ các thuộc tính đã biết. Sau khi mô hình được tạo, nó được sử dụng để xử lý dữ liệu có cấu trúc giống dữ liệu đầu vào. Trong học không giám sát, dữ liệu đầu vào chưa có nhãn, không có cấu trúc. Nhiệm vụ của thuật toán là xác định một cấu trúc trong dữ liệu.[2].

Phương pháp phân tích thành phần chính (Principle Component Analysis – PCA) là một kết quả rất đẹp và quan trọng của đại số tuyến tính. Ngày nay kết quả này được ứng dụng trong rất nhiều lĩnh vực như: Công nghệ thông tin, Học máy, sinh học, tài chính.[1], [2].

Đối với các lĩnh vực ứng dụng sử dụng dữ liệu, chẳng hạn như học máy, dữ liệu cần phân tích ban đầu phụ thuộc nhiều biến, vấn đề là các biến này thường có tương quan với nhau sẽ bất lợi cho việc áp dụng các biến này để xây dựng các mô hình tính toán ví dụ: Hồi quy... và với số biến giải thích lớn chúng ta sẽ rất khó để có cái nhìn trực quan về dữ liệu ví dụ: thị trường ta quan tâm có hàng ngàn mã cổ phiếu làm cách nào để khi quan sát dữ liệu từ hàng ngàn cổ phiếu này ta hình dung được xu hướng của toàn thị trường.

Phương pháp PCA sẽ "chiếu" (biểu diễn) dữ liệu đa chiều lên một không gian có cơ sở trực giao, tức nếu ta xem mỗi cơ sở trong không gian mới là

một biến thì hình ảnh của dữ liệu gốc trong không gian mới này sẽ được biểu diễn thông qua các biến độc lập (tuyến tính). Vấn đề: nếu chuyển dữ liệu ban đầu sang không gian mới thì những thông tin đáng quan tâm của dữ liệu ban đầu liệu có bị mất? Để giải quyết vấn đề này phương pháp PCA sẽ tìm không gian mới với tiêu chí cố gắng phản ánh được càng nhiều thông tin gốc càng tốt, và thước đo cho khái niệm "thông tin" ở đây là phương sai. Một điểm hay nữa là: do các biến trong không gian mới độc lập, nên ta có thể tính toán được tỷ lệ giải thích phương sai của từng biến mới đối với dữ liệu, điều này cho phép ta cân nhắc việc chỉ dùng số ít các biến để giải thích dữ liệu.

Các ứng dụng tự nhiên mà ta có thể nhận ra là:

- Giảm kích thước của dữ liệu
- Nếu ta có thể giảm số chiều về 2 hoặc 3 chiều, ta có thể dùng các loại đồ thị để hiểu thêm về dữ liệu mà mình đang có, giúp ta nhìn dữ liệu trực quan hơn.
- Xử lý vấn đề tương quan giữa các biến trong dữ liệu ban đầu bằng cách sử dụng các biến mới trong không gian mà phương pháp PCA tìm được để mô tả dữ liệu.
- Nén ảnh: Giảm kích thước ảnh mà vẫn có thể giữ được những đặc trưng quan trọng.

Đối tượng và phạm vi nghiên cứu Nghiên cứu các kiến thức về đại số tuyến tính liên quan PCA và bản thân phương pháp giảm chiều dữ liệu sử dụng PCA. Phạm vi nghiên cứu: Nghiên cứu kiến thức nền tảng của PCA để hiểu bản chất; Nghiên cứu ứng dụng PCA trong học máy để giảm chiều dữ liệu trong các trường hợp.

Nội dung của luận văn gồm 3 chương.

Chương 1. Tổng quan học máy và bài toán giảm chiều dữ liệu

Chương này trình bày các kiến thức tổng quan về học máy và bài toán

giảm chiều dữ liệu cũng như một số nền tảng toán học cần thiết. Nội dung bao gồm

1.1 Tổng quan về học máy.

1.2 Tổng quan về giảm chiều dữ liệu.

1.3 Nền tảng toán học.

Chương 2. Phương pháp PCA giảm chiều dữ liệu

Nội dung chương 2 tập trung nghiên cứu thuật toán PCA, bao gồm các mục:

2.1 Phát biểu bài toán

2.2 Phân tích thành phần chính

Chương 3. Một số ứng dụng của PCA.

Chương này nghiên cứu, cài đặt một số ứng dụng của PCA.Nội dung bao gồm 3.1 Khuôn mặt riêng.

3.2 Dò tìm điểm bất thường.

3.3 Ứng dụng PCA trong tài chính.

3.4 Ứng dụng PCA trong trực quan hóa dữ liệu, khử nhiễu.

Cuối cùng là phần kết luận: Trình bày kết quả mà luận văn đạt được và hướng phát triển.

Mặc dù đã có cố gắng nỗ lực, song luận văn không tránh khỏi những thiếu sót do năng lực và thời gian hạn chế. Em chân thành mong muốn lắng nghe những đóng góp, góp ý của thầy cô bạn bè đồng nghiệp để luận văn được cải thiện tốt hơn.

Em xin chân thành cảm ơn.