ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

GIANG THÀNH TRUNG

NGHIÊN CỨU MỘT SỐ PHƯƠNG PHÁP GIẢM SỐ CHIỀU DỮ LIỆU

Chuyên ngành: Hệ thống thông tin
Mã số: 9480104.01

TÓM TẮT LUẬN ÁN TIẾN SĨ HỆ THỐNG THÔNG TIN

Hà Nội - 2021

Công trình được hoàn thành tại: Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội.

Người hướng dẫn khoa học: **- PGS. TS. Trần Đăng Hưng**
- TS. Lê Nguyên Khôi

Phản biện: .....

.....

Phản biện: .....

.....

Phản biện: .....

.....

Luận án sẽ được bảo vệ trước Hội đồng cấp Đại học Quốc gia chấm luận án tiến sĩ họp tại ..... vào hồi ..... giờ ..... ngày ..... tháng ..... năm .....

Có thể tìm hiểu luận án tại:
- Thư viện Quốc gia Việt Nam.
- Trung tâm Thông tin - Thư viện, Đại học Quốc gia Hà Nội

MỞ ĐẦU

Đặt vấn đề

Trong thập kỷ vừa qua, ngành khoa học đời sống và thực nghiệm đã trải qua một cuộc cách mạng với sự phát triển nhanh chóng của các thiết bị thí nghiệm và thiết bị đo công nghệ cao. Cùng với sự phát triển đó, lượng dữ liệu được đo đạc, lưu trữ và xử lý ngày càng lớn trên tất cả các lĩnh vực của đời sống xã hội, đặc biệt trong lĩnh vực y sinh học đã có sự phát triển vượt bậc về dữ liệu kể từ khi bộ trình tự gene hoàn chỉnh của con người được giải mã. Nhiều bộ dữ liệu y sinh học có sự gia tăng theo hàm mũ và thường tồn tại ở nhiều dạng khác nhau như: vector số, ảnh, âm thanh, video, văn bản,... Nguồn dữ liệu này là cơ sở cho việc phân tích và đề xuất trong các hệ thống trợ giúp ra quyết định hỗ trợ cho các hoạt động chuẩn đoán và chữa trị các bệnh do chúng chính là thông tin phản ánh khách quan các hoạt động đã xảy ra trong chính các cơ quan của cơ thể.

Dữ liệu ở dạng thô được xử lý, biến đổi, tính toán và chuyển hóa thành tri thức để trở nên hữu ích nhằm hỗ trợ ra quyết định. Tuy nhiên, một trong các thách thức đối với các phương pháp xử lý dữ liệu đó là sự mất cân bằng giữa số lượng thuộc tính (còn gọi là đặc trưng, biến) và số lượng mẫu quan sát. Trong nhiều trường hợp, các bộ dữ liệu có số thuộc tính nhiều hơn rất nhiều so với số lượng đối tượng quan sát được (mẫu). Ví dụ, một tập dữ liệu microarray là một mảng hai chiều, trong đó mỗi cột là một gen, mỗi dòng là một mẫu quan sát. Đối với mỗi loại sinh vật, số lượng gen thường từ vài nghìn đến vài chục nghìn, trong khi đó số mẫu chỉ thường vài trăm. Nếu nhìn theo khía cạnh hệ phương trình toán học, đó là bài toán có số phương trình ít hơn rất nhiều lần so với số biến và đó là bài toán không giải được trong thời gian đa thức. Ngoài ra, khi tập dữ liệu ngày càng lớn kèm theo đó là số lượng biến lớn thì đòi hỏi chi phí tính toán lâu, dẫn đến không đáp ứng được nhu cầu về thời gian phản hồi khi đưa vào trong các bài toán thực tế. Khi đó, một bước tiền xử lý được đề xuất là giảm chiều dữ liệu nhằm giảm số lượng biến để phù hợp với các hệ thống máy tính và các mô hình tính toán ở bước tiếp theo.

Giảm chiều dữ liệu được hiểu là từ một tập dữ liệu gốc ban đầu, áp dụng các phương pháp phân tích để giảm rất nhiều đặc trưng của dữ liệu sao cho vẫn giữ lại được bản chất thông tin của tập dữ liệu đó. Giảm chiều dữ liệu hiện nay đã trở thành một bước kỹ thuật cần thiết nhằm biến đổi dữ liệu gốc ban đầu bằng cách giảm đặc trưng để phù hợp với số mẫu và các mô hình tính toán ở bước tiếp theo. Trong nhiều năm qua, hướng nghiên cứu về giảm chiều dữ liệu luôn thu hút được sự quan tâm của các nhà nghiên cứu và thực tế đã có rất nhiều phương pháp giảm chiều dữ liệu đã được đưa ra nhằm giải quyết bài toán nêu trên. Trong lĩnh vực Tin-sinh học, giảm chiều dữ liệu đã được ứng dụng rộng rãi vào trong một số kỹ thuật xử lý của các bài toán như: giảm chiều dữ liệu các tập dữ liệu sinh học phân tử đơn lẻ; sử dụng các phương pháp giảm chiều dữ liệu để trích rút các thông tin hữu ích trong các tập dữ liệu sinh học phân tử; kết hợp đồng thời giảm chiều dữ liệu và tích hợp các tập dữ liệu sinh học phân tử,... Tuy nhiên, cùng với sự phát triển của ngành khoa học dữ liệu, các tập dữ liệu hiện nay trở nên rất đa dạng, có cấu trúc và mối quan hệ phức tạp, đặc biệt là có kích thước lớn và được biểu diễn bởi nhiều độ đo khác nhau. Do đó, các nghiên cứu giảm chiều dữ liệu cũng phải đối mặt với các thách thức mới xuất hiện, bao gồm: Một là, các tập dữ liệu gồm nhiều dữ liệu nhiễu, thưa và ngoại lai, nếu phân tích chung với dữ liệu thông thường sẽ cho ra kết quả không chính xác; Hai là, các loại dữ liệu

sinh học phân tử khác nhau đều chứa những thông tin hữu ích về các con đường phân tử trong tế bào và vai trò của chúng đối với bệnh tật, do đó một nhu cầu bức thiết là tích hợp các nguồn dữ liệu có ý nghĩa đồng thời với giảm chiều dữ liệu để tạo ra tập dữ liệu tích hợp mang đầy đủ thông tin nhưng vẫn phù hợp với các mô hình, công cụ tính toán hiện tại. Tuy nhiên, bản thân mỗi loại dữ liệu đã có kích thước lớn, ngoài ra, mỗi loại dữ liệu sử dụng những độ đo khác nhau, do đó, việc tích hợp dữ liệu cũng còn tồn tại nhiều thách thức.

Từ những phân tích nêu trên, tác giả chọn thực hiện luận án **Nghiên cứu một số phương pháp giảm số chiều dữ liệu** làm đề tài nghiên cứu tiến sĩ của mình. Thông qua nghiên cứu này, luận án tập trung vào giải quyết một số vấn đề lớn sau:

Thứ nhất, nghiên cứu về các phương pháp giảm chiều dữ liệu đã được đề xuất, xác định ưu, nhược điểm của các phương pháp đã được đề xuất, từ đó làm tiền đề đề xuất những cải tiến nhằm nâng cao hiệu quả của các phương pháp giảm chiều dữ liệu.

Thứ hai, nghiên cứu cụ thể về một số phương pháp có tính ứng dụng cao, phân tích ưu, nhược điểm của phương pháp để từ đó đề xuất cải tiến nhằm nâng cao hiệu quả của phương pháp.

Thứ ba, áp dụng các phương pháp đã nghiên cứu vào xây dựng các mô hình phân lớp bệnh nhân để khẳng định tính ứng dụng cũng như hiệu quả của phương pháp. Đặc biệt chú ý đến việc tích hợp dữ liệu từ nhiều nguồn khác nhau nhằm tận dụng sự phong phú của các nguồn dữ liệu cũng như thông tin hữu ích trong đó.

Mục tiêu của luận án

Tác giả đặt ra ba mục tiêu lớn cần đạt được của luận án dựa trên các vấn đề cần giải quyết như sau:

1. Nghiên cứu và tổng hợp để xây dựng tổng quan về các phương pháp giảm chiều dữ liệu, tập trung vào các phương pháp được ứng dụng trong lĩnh vực Tin-Sinh học. Trong đó, tập trung thảo luận ưu, nhược điểm của các phương pháp đã được đề xuất.
2. Phân tích hai phương pháp hiệu quả trong xử lý dữ liệu Tin-Sinh học là Phương pháp học đa nhân kết hợp giảm chiều dữ liệu và Phương pháp phân tích thành phần chính tăng cường. Trên cơ sở đó đó tìm ra những điểm mạnh và hạn chế của các phương pháp đó để đề xuất một số cải tiến nhằm khắc phục những tồn tại đã chỉ ra để tăng tính ứng dụng của các phương pháp trên trong thực tế.
3. Căn cứ trên các đề xuất cải tiến, xây dựng mô hình phân lớp bệnh nhân nhằm tạo ra một công cụ hỗ trợ quá trình chuẩn đoán và điều trị bệnh. Các mô hình được đề xuất là ứng dụng thực tiễn của kết quả nghiên cứu lý thuyết đạt được ở mục tiêu thứ hai.

Các đóng góp chính của luận án

Luận án sau khi được thực hiện đã có ba đóng góp chính sau:

1. Đề xuất một thuật toán hiệu quả dựa trên học đa nhân kết hợp giảm chiều dữ liệu (Phương pháp fMKL-DR). Xuất phát từ những tồn tại của phương pháp Học đa nhân kết hợp giảm chiều dữ liệu (MKL-DR - Một phương pháp phù hợp và hiệu quả trong tiền xử lý dữ liệu y sinh) là đòi hỏi chi phí về mặt thời gian lớn do trong thuật toán lặp đi lặp lại việc tính tích chuỗi ma trận. Tác giả đã đề xuất một thuật toán dựa trên phương pháp quy hoạch động để xác định thứ tự nhân tối ưu cho tích chuỗi ma trận. Từ đó, tác giả đã đề xuất một phương pháp cải tiến của phương pháp MKL-DR gọi là fMKL-DR nhằm giảm đáng kể chi phí về thời gian tính toán mà vẫn giữ được hiệu quả của phương pháp. Kết quả này có ý nghĩa rất lớn trong thời đại dữ liệu lớn

hiện nay khi tập dữ liệu ngày càng lớn, đa dạng và việc giảm đáng kể thời gian thực hiện của phương pháp sẽ giúp tăng khả năng ứng dụng của phương pháp trong thực tế. Kết quả này được công bố trong bài báo [GTTrung-1] tại hội thảo KSE 2017 và [GTTrung-2] tại Hội thảo IUKM 2018.

2. Dựa trên phương pháp fMKL-DR đã đề xuất ở trên, tác giả đề xuất mô hình phân lớp bệnh nhân gồm:

   - Mô hình tích hợp dữ liệu bệnh nhân ung thư từ các nguồn dữ liệu khác nhau và thực hiện phân tầng bệnh nhân ung thư hiệu quả. Mô hình được đề xuất đã kết hợp dữ liệu dựa trên fMKL-DR từ ba loại dữ liệu khác nhau gồm: biểu hiện gene, methyl hóa DNA và biểu hiện miRNA hoặc biểu hiện Protein. Mô hình phân lớp dựa trên phương pháp Máy vector hỗ trợ (SVM) với đầu vào là tập dữ liệu đã được tích hợp bởi fMKL-DR. Mô hình đã có kết quả phân tách tốt, kết quả này đóng góp vào việc hỗ trợ, chuẩn đoán trong điều trị bệnh ung thư.
   - Mô hình phân lớp bệnh nhân Alzheimer dựa trên dữ liệu ảnh cộng hưởng từ. Dữ liệu ảnh cộng hưởng từ dưới dạng ảnh sẽ được phân tích bằng phần mềm FreeSurfer, sau đó, trích xuất 6 giá trị độ đo được đánh giá là hiệu quả trong việc phân tích dữ liệu bệnh Alzheimer từ ảnh chụp cộng hưởng từ và thu được 6 tập dữ liệu tương ứng. Một mô hình tích hợp dữ liệu từ 6 tập dữ liệu dựa trên phương pháp fMKL-DR và thực hiện xây dựng mô hình phân lớp dựa trên SVM. Mô hình được xây dựng đã có kết quả phân tách rất tốt, trợ giúp cho quá trình phát hiện sớm và đúng trạng thái bệnh của người bệnh để có phác đồ điều trị bệnh phù hợp.

   Các mô hình phân lớp bệnh nhân được đề xuất đã có kết quả tích cực và là công cụ hiệu quả hỗ trợ trong điều trị bệnh ung thư và bệnh Alzheimer. Các mô hình này đã tận dụng được thế mạnh của fMKL-DR là có thể tích hợp nhiều nguồn dữ liệu khác nhau đồng thời với giảm chiều dữ liệu mà vẫn bảo đảm hiệu năng về mặt thời gian tính toán. Mô hình này có tính ứng dụng cao khi dữ liệu y sinh được quan sát, lưu trữ và đưa vào xử lý ngày càng đa dạng về loại hình cũng như độ phức tạp về kích thước (xét ở khía cạnh số đặc trưng). Kết quả này đã được công bố trong bài báo [GTTrung-3] trên Tạp chí BMC Medical Informatics and Decision Making năm 2020.

3. Đề xuất mô hình phân lớp bệnh nhân ung thư dựa trên phương pháp Phân tích thành phần chính tăng cường (RPCA). Trong đó đề xuất một hướng giảm chiều dữ liệu bằng cách lựa chọn các đặc trưng dựa trên RPCA phù hợp với tập dữ liệu Tin-sinh học. Từ đó làm căn cứ xây dựng mô hình phân lớp bệnh nhân.

   Kết quả này được công bố trên bài báo [GTTrung-4] tại Hội thảo AICI 2021.

Đóng góp 1, 2 được trình bày trong nội dung của Chương 2, đóng góp 3 được trình bày trong nội dung của Chương 3. Ngoài các đóng góp trên, luận án còn trình bày các nội dung kiến thức khác phụ trợ cho các phương pháp chính được trình bày trong mỗi chương.

Bố cục của luận án

Bố cục của luận án gồm 5 phần chính:

- **Mở đầu** trình bày khái quát về bài toán giảm chiều dữ liệu và ứng dụng trong lĩnh vực Tin-sinh học. Ngoài ra, phần này cũng trình bày về các đóng góp chính của luận án và bố cục của luận án.
- **Chương 1, 2, 3** là phần nội dung của luận án tương ứng với 3 nội dung cụ thể sau:

  Chương 1 trình bày ***Tổng quan về giảm chiều dữ liệu và ứng dụng trong xử lý dữ liệu Tin-sinh học***

Chương 2 trình bày một phương pháp giảm chiều dữ liệu được ứng dụng hiệu quả trong bài toán Tin-sinh học là phương pháp ***Phân lớp bệnh nhân hiệu quả dựa trên học đa nhân kết hợp giảm chiều dữ liệu***

Chương 3 trình bày phương pháp ***Phân lớp bệnh nhân dựa trên phương pháp phân tích thành phần chính tăng cường***

- **Kết luận** tóm lược lại các kết quả đã đạt được của luận án, từ đó phân tích những hạn chế và các hướng nghiên cứu tiếp theo phù hợp với nội dung của luận án trong tương lai.

Chương 1

TỔNG QUAN VỀ GIẢM CHIỀU DỮ LIỆU VÀ ỨNG DỤNG TRONG TIN-SINH HỌC

Chương này, tập trung trình bày tổng quan về giảm chiều dữ liệu, tầm quan trọng cũng như ứng dụng của các phương pháp giảm chiều trong xử lý dữ liệu Tin-sinh học nhằm đưa ra bức tranh tổng quan về các phương pháp giảm chiều dữ liệu.

1.1. Bài toán giảm chiều dữ liệu

Bài toán giảm chiều dữ liệu được phát biểu như sau:
Đầu vào: Tập dữ liệu X có số chiều (đặc trưng, biến) S lớn
Đầu ra: Tập dữ liệu X' có số chiều D nhỏ hơn rất nhiều so với S mà vẫn giữ được bản chất dữ liệu tương đương X.
Bài toán giảm chiều dữ liệu là bài toán đi tìm một hàm số:

$$\begin{aligned} f :&\mathbb{R}^S \to \mathbb{R}^D \\ &x \to z \end{aligned} \tag{1.1}$$

với $S < D$, hàm f biến một điểm dữ liệu x trong không gian có số chiều (đặc trưng) lớn $\mathbb{R}^S$ thành một điểm z trong không gian có số chiều nhỏ $\mathbb{R}^D$.

1.2. Các hướng tiếp cận trong nghiên cứu giảm chiều dữ liệu

1.2.1. Lựa chọn đặc trưng

1.2.1.1. Các phương pháp lọc (Filter Methods)

1.2.1.2. Các phương pháp bao gói (Wrapper Methods)

1.2.1.3. Các phương pháp nhúng (Embedded Methods)

1.2.2. Trích chọn đặc trưng

1.2.2.1. Phương pháp trích chọn đặc trưng tuyến tính

1.2.2.2. Phương pháp trích chọn đặc trưng không tuyến tính

1.2.3. Phương pháp lai

1.3. Ý nghĩa và ứng dụng của giảm chiều dữ liệu

Giảm chiều dữ liệu được áp dụng thực tế trong nhiều lĩnh vực như:

1. Xử lý ảnh.
2. Xử lý ngôn ngữ tự nhiên.
3. Các bài toán trong Tin-sinh học.
4. Một số lĩnh vực khác.

1.4. Kết luận

Giảm chiều dữ liệu đã trở thành một bước tiền xử lý đóng vai trò quan trọng trong quá trình Khai phá tri thức từ dữ liệu ở nhiều lĩnh vực. Dữ liệu được biến đổi từ không gian có số chiều cao (với nhiều tồn tại như không phù hợp với mô hình tính toán, chứa nhiều nhiễu, dữ liệu thưa) sang không gian có số chiều thấp hơn (phù hợp với mô hình tính toán, loại bỏ nhiễu, cô đặc dữ liệu). Đã có rất nhiều phương pháp giảm chiều dữ liệu được đề xuất thuộc một trong ba nhóm phương pháp (lựa chọn đặc trưng, trích chọn đặc trưng, lai giữa hai phương pháp trên) và đã chứng minh được hiệu quả của chúng. Tuy nhiên, từ những phân tích, đánh giá các phương pháp ở trên tác giả nhận thấy vẫn còn tồn tại một số thách thức đối với bài toán giảm chiều dữ liệu mà các phương pháp được đề xuất vẫn chưa khắc phục được triệt để như:

Một là, các phương pháp hiện nay thường đòi hỏi chi phí tính toán lớn. Một số phương pháp lưu trữ các trạng thái để huấn luyện mô hình và đòi hỏi chi phí về bộ nhớ lớn. Ngoài ra, dữ liệu được biểu diễn dưới dạng ma trận và khi thực hiện các phép toán tính tích để tổ hợp ma trận thì thường đòi hỏi chi phí rất lớn về mặt thời gian. Đây là một trong những thách thức không nhỏ khi thực tế lượng dữ liệu ngày càng tăng và để phù hợp áp dụng trong thực tế thì tốc độ đáp ứng về mặt thời gian cần phải được đảm bảo.

Hai là, các phương pháp đã được đề xuất thường sử dụng khá nhiều tham số trong mô hình. Việc sử dụng tham số giúp phương pháp có sự linh động trong việc sử dụng trong nhiều bài toán khác nhau. Tuy nhiên, với mỗi bài toán cụ thể, việc tìm ra bộ tham số tối ưu cũng mất khá nhiều thời gian của các nhà nghiên cứu khi làm thực nghiệm do mỗi bộ tham số chỉ phù hợp với những đặc trưng dữ liệu nhất định. Từ đó cho thấy, cần có một giải pháp nghiên cứu, đề xuất cách lựa chọn tham số phù hợp với bài toán, loại dữ liệu giúp giảm thời gian làm những thực nghiệm không mang nhiều ý nghĩa của các nhà nghiên cứu.

Ba là, hầu hết các phương pháp được đề xuất thường dựa trên một bài toán ứng dụng cụ thể nên thường có kết quả rất tốt khi áp dụng vào bài toán cụ thể đó. Tuy nhiên, khi áp dụng các phương pháp đó sang các bài toán khác thì thường không đạt được kết quả tốt, nếu có cần chỉnh sửa hoặc bổ sung thêm nhiều các thành phần để sử dụng trên bài toán mới. Đây cũng là một điều rất đáng tiếc, nếu có một phương pháp giảm chiều dữ liệu có mức tổng quát cao, bao hàm được phạm vi rộng rãi các bài toán, các loại dữ liệu thì sẽ có ý nghĩa rất lớn.

Bốn là, đối với một số bài toán có sự tương đồng dữ liệu cao, khi áp dụng các phương pháp giảm chiều dữ liệu thì tập dữ liệu sau khi giảm chiều có sự phân tách chưa thực sự tốt. Có thể kể đến bài toán chuẩn đoán xem bệnh nhân có triệu chứng suy giảm nhận thức nhẹ sẽ bị chuyển sang bệnh Alzheimer hay không cũng chưa đạt được kết quả tốt do các bệnh nhân đều có các đặc trưng tương đồng nhau (xét trên ảnh chụp cộng hưởng từ não).

Từ bốn thách thức nêu trên cho thấy bài toán giảm chiều dữ liệu vẫn là một hướng nghiên cứu hấp dẫn, thu hút được sự quan tâm của các nhà nghiên cứu. Mỗi phương pháp mới được đề xuất, mỗi cải tiến hoặc đề xuất ứng dụng sẽ góp phần hỗ trợ cho việc phát hiện các tri thức hữu ích từ dữ liệu.

Chương 2

PHƯƠNG PHÁP HIỆU QUẢ PHÂN LỚP BỆNH NHÂN KẾT HỢP GIẢM CHIỀU DỮ LIỆU

Chương này trình bày phương pháp hiệu quả dựa trên giảm chiều dữ liệu kết hợp học đa nhân và đề xuất mô hình phân lớp bệnh nhân dựa trên phương pháp đã đề xuất. Cụ thể, tác giả đề xuất phương pháp hiệu quả dựa trên học đa nhân kết hợp giảm chiều dữ liệu (fMKL-DR). fMKL-DR dựa trên tối ưu công thức tính toán tích chuỗi ma trận thông qua một thuật toán xác định thứ tự nhân tích chuỗi ma trận từ đó làm giảm đáng kể thời gian tính toán của phương pháp. Ngoài ra, dựa trên phương pháp fMKL-DR, tác giả đề xuất một mô hình hiệu quả để phân lớp bệnh nhân ung thư và phân lớp bệnh nhân Alzheimer. Mô hình phân lớp được đề xuất là một công cụ hiệu quả làm tiền đề ứng dụng hỗ trợ trong việc phát hiện và điều trị các bệnh nói trên. Các kết quả của Chương này đã được công bố trong các bài báo [GTTrung-1], [GTTrung-2] và [GTTrung-4].

2.1. Giới thiệu

2.2. Kiến thức nền tảng

2.2.1. Phương pháp nhân

2.2.2. Phương pháp học đa nhân

2.2.3. Phương pháp nhúng đồ thị

2.3. Phương pháp MKL-DR

2.3.1. Ý tưởng thuật toán

Giảm chiều dữ liệu kết hợp học đa nhân (Multiple Kernel Learning and Dimensionality Reduction - MKL-DR) được đề xuất bởi Lin và cộng sự. Phương pháp kết hợp cả học đa nhân và giảm chiều dữ liệu dựa trên nhúng đồ thị nhằm vừa tích hợp dữ liệu đồng thời giảm chiều dữ liệu. Bài toán MKL-DR trong không gian nhiều chiều được phát biểu như sau:

$$\min_{A,\beta} \sum_{i,j=1}^{N} \left|\left|A^\top \mathbb{K}^{(i)}\beta - A^\top \mathbb{K}^{(j)}\beta\right|\right|^2 w_{ij} \tag{2.7}$$

$$\text{s.t.} \sum_{i,j=1}^{N} \left|\left|A^\top \mathbb{K}^{(i)}\beta - A^\top \mathbb{K}^{(j)}\beta\right|\right|^2 w_{ij} = \text{const};$$

$$\beta_m \geq 0, m = 1, ..M.$$

Bài toán (2.7) cần được tối ưu dựa trên cả A và β. Việc giải bài toán tối ưu dựa trên đồng thời cả 2 biến là rất khó, vì vậy, một giải pháp được sử dụng là tối ưu hóa trên từng biến một, nghĩa là, tại mỗi lần lặp thì A hoặc β sẽ được tối ưu trong khi biến còn lại sẽ được cố định, sau đó thực hiện ngược lại.

Cố định A để tìm β. Bài toán (2.7) trở thành bài toán tối ưu bậc 2 dựa trên các ràng buộc bậc 2, bài toán này nằm trong lớp bài toán NP-Khó, để giải được có thể rút gọn thành bài toán semidefinite và giải bằng semidefinite programming trong như sau:

$$\min_{\beta, B} \text{trace}(S_W^A B) \tag{2.8}$$

$$\text{s.t. trace}(S_D^A B) = 1 \quad \text{hoặc} \quad \text{trace}(S_{W'}^A B) = 1;$$

$$e_m^\top \beta \geq 0, m = 1, ..., M;$$

$$\begin{bmatrix} 1 & \beta^\top \\ \beta & B \end{bmatrix} \succeq 0.$$

với:

$$S_W^A = \sum_{i,j=1}^{N} w_{ij} (\mathbb{K}^{(i)} - \mathbb{K}^{(j)})^\top A A^\top (\mathbb{K}^{(i)} - \mathbb{K}^{(j)}) \tag{2.9}$$

$$S_{W'}^A = \sum_{i,j=1}^{N} w'_{ij} (\mathbb{K}^{(i)} - \mathbb{K}^{(j)})^\top A A^\top (\mathbb{K}^{(i)} - \mathbb{K}^{(j)}) \tag{2.10}$$

với e_m là các vector cột mà tất cả các phần tử bằng 0 ngoại trừ phần tử thứ m bằng 1 và B là biến được thêm vào để rút gọn bài toán gốc về bài toán semidefinite-relaxation.

Cố định β để tìm A. Bài toán (2.7) trở thành bài toán giá trị riêng tổng quát sau:

$$S_W^\beta \alpha = \lambda S_D^\beta \alpha \text{ hoặc } S_W^\beta \alpha = \lambda S_{W'}^\beta \alpha \tag{2.11}$$

với:

$$S_W^\beta = \sum_{i,j=1}^{N} w_{ij} (\mathbb{K}^{(i)} - \mathbb{K}^{(j)}) \beta \beta^\top (\mathbb{K}^{(i)} - \mathbb{K}^{(j)})^\top \tag{2.12}$$

$$S_{W'}^\beta = \sum_{i,j=1}^{N} w'_{ij} (\mathbb{K}^{(i)} - \mathbb{K}^{(j)}) \beta \beta^\top (\mathbb{K}^{(i)} - \mathbb{K}^{(j)})^\top \tag{2.13}$$

Việc lặp sẽ thực hiện cho đến khi hội tụ hoặc đạt tối đa số lần lặp. Có thể khởi tạo giá trị ban đầu cho A khi tối ưu β trước bằng cách gán cho $A^\top A = I$, hoặc khởi tạo vector trọng số β với các phần tử bằng nhau và bằng 1 nếu tối ưu A trước.

Thuật toán 1 dưới đây mô tả thủ tục huấn luyện MKL-DR. Đầu vào của thuật toán chính là M tập dữ liệu, mỗi tập dữ liệu sẽ được biểu diễn thành một ma trận nhân (gọi là ma trận nhân cơ sở) và một phương pháp giảm chiều dữ liệu được xác định dựa trên các ma trận W và W'.

Thời gian thực hiện thuật toán có thể chia thành hai giai đoạn: tích hợp dữ liệu kết hợp giảm chiều dữ liệu và xây dựng mô hình phân lớp dữ liệu. Tích hợp dữ liệu kết hợp

Thuật toán 1: Thuật toán MKL-DR

input : Phương pháp giảm chiều được xác định dựa trên các ma trận W và W' theo (2.3);
M ma trận nhân cơ sở ${K_m}_{m=1}^{M}$ tương ứng với M tập dữ liệu.

output: Ma trận hệ số mẫu (ma trận chiếu) $A = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, ..., \boldsymbol{\alpha}_P]$;
Vector trọng số nhân β.

1 **begin**
2 Khởi tạo giá trị cho A hoặc β;
3 **for** $t \leftarrow 1, 2, ..., T$ **do**
4 Tính S_W^A dựa trên (2.9) và $S_{W'}^A$ dựa trên (2.10);
5 Giải bài toán tối ưu (2.8) bằng SDP để tìm β;
6 Tính S_W^β dựa trên (2.12) và $S_{W'}^\beta$ dựa trên (2.13);
7 Giải bài toán giá trị riêng tổng quát (2.11) để tìm A;
8 **return** A, β;

giảm chiều dữ liệu được thực hiện bằng cách lặp và cập nhật lại ma trận chiếu A và vector trọng số nhân β. Việc giải tìm ra β bằng cách giải bài toán semidefinite programming với số các ràng buộc tuyến tính với số ma trận nhân đầu vào và số biến sẽ bậc hai với số ma trận nhân đầu vào. Tuy nhiên, nếu $M \ll N$, thì thời gian thực hiện không đáng kể, khi đó thời gian sẽ tập trung vào việc giải bài toán giá trị riêng để tìm A. Khi đó độ phức tạp của thuật toán là $\mathcal{O}^3$.

2.3.2. Nhận xét phương pháp MKL-DR

*** Ưu điểm**
*** Hạn chế**

Thuật toán 1 đã thể hiện thuật toán MKL-DR. Phần lớn thời gian tính toán của thủ tục huấn luyện MKL-DR là để tính các giá trị $S_W^A, S_{W'}^A, S_W^\beta, S_{W'}^\beta$ (tại dòng 4, 6 của Thuật toán 1). Các giá trị được tính toán bởi các công thức (2.9), (2.10), (2.12), (2.13) tương ứng và được tính toán lặp lại T lần. Mỗi công thức này lại là tích chuỗi các ma trận, độ phức tạp của tích chuỗi ma trận phụ thuộc vào kích thước của các ma trận thành phần (số mẫu N). Do vậy, nếu kích thước số mẫu quan sát càng lớn thì chi phí tính toán cũng gia tăng đáng kể. Trong một số thực nghiệm, với 541 mẫu quan sát, MKL-DR cần tới hàng nghìn giây để huấn luyện. Hạn chế này đã vi phạm một trong ba nguyên tắc trong đánh giá một phương pháp phân tích dữ liệu. Chính vì vậy, cải thiện hiệu suất về mặt thời gian tính toán cho MKL-DR là rất cần thiết để phương pháp này có thể áp dụng trong thực tế.

2.4. Đề xuất cải tiến phương pháp MKL-DR

Sau khi phân tích, có ba tham số ảnh hưởng đến hiệu năng của phương pháp MKL-DR bao gồm: số lượng mẫu quan sát N, số loại dữ liệu M và số chiều sau khi đã giảm P. Thông thường, M có giá trị nhỏ và nằm trong đoạn từ 3-10, số chiều sau khi giảm cũng nhỏ (trong thực nghiệm, tác giả chọn $P = 5$). Do đó độ phức tạp của thuật toán là $\mathcal{O}(N^3)$ với thời gian đa thức. Thuật toán huấn luyện MKL-DR tính toán lặp lại nhiều lần các công thức (2.9), (2.10), (2.12), (2.13), các công thức này tính dựa trên tổng xích ma các tích chuỗi ma trận. Khi các ma trận thành phần có kích thước càng lớn thì độ phức tạp của thuật toán càng lớn. Do vậy, thời gian huấn luyện của MKL-DR sẽ gia tăng nhanh chóng khi số lượng các mẫu quan sát tăng.

Độ phức tạp của tích chuỗi các ma trận chính là số hoạt động nhân cần thực hiện khi nhân từng cặp ma trận. Tích chuỗi ma trận có thuộc tính tổ hợp, nghĩa là khi thay đổi

thứ tự nhân giữa các cặp ma trận thì số lượng phép nhân sẽ thay đổi mà không ảnh hưởng đến kết quả của phép nhân. Do đó, tác giả đề xuất một thủ tục dựa trên quy hoạch động để tìm thứ tự thực hiện các phép nhân sao cho số phép nhân phải thực hiện là nhỏ nhất. Nếu số lượng các phép nhân là nhỏ nhất thì thời gian tính toán công thức sẽ giảm giúp tăng hiệu suất về mặt thời gian của thuật toán.

2.4.1. Thuật toán tìm thứ tự tính toán tốt nhất cho tích chuỗi ma trận

Phát biểu bài toán: Cho N ma trận $A_1, A_2, ..., A_N$ với kích thước của ma trận A_i là $d_{i-1} \times d_i$. Tìm thứ tự nhân các ma trận $A_1 \times A_2 \times ... \times A_N$ sao cho số phép nhân phải thực hiện là nhỏ nhất.

Tác giả phát triển một thuật toán để cải tiến hoạt động nhân của tích chuỗi ma trận sao cho tối thiểu hóa số phép nhân cần thực hiện trong Thuật toán 2 dưới đây.

Thuật toán 2: Thuật toán tìm thứ tự tính tích chuỗi ma trận (MCMO)

input : N ma trận kích thước $d_1, d_2, ..., d_N$.

output: Thứ tự nhân $O = [o_1, o_2, ..., o_q]$ sao cho số phép nhân cần thực hiện là nhỏ nhất.

begin
- $F(i,i) = 0, i = 1, ..., N$;
- $F(i, i+1) \leftarrow d_{i-1} \times d_i \times d_{i+1}$;
- $O = [\,]$;
- **for** $i \leftarrow 1, 2, ..., N-1$ **do**
  - **for** $j \leftarrow i+1, i+2, ..., N$ **do**
    - $F(i,j) = min(F(i,t) + F(t+1,j) + d_{i-1} \times d_t \times d_j)$
    - Thêm t vào O;
- **return** O;

Thuật toán 3: Thuật toán hiệu quả dựa trên giảm chiều dữ liệu kết hợp học đa nhân (fMKL-DR)

begin
- Khởi tạo giá trị cho A hoặc β;
- $O_A = \text{MCMO}(\text{size}(\mathbb{K}^{(i)} - \mathbb{K}^{(j)}), \text{size}(\beta), \text{size}(\beta^\top), \text{size}((\mathbb{K}^{(i)} - \mathbb{K}^{(j)})^\top))$;
- $O_\beta = \text{MCMO}(\text{size}((\mathbb{K}^{(i)} - \mathbb{K}^{(j)})^\top), \text{size}(A), \text{size}(A^\top), \text{size}(\mathbb{K}^{(i)} - \mathbb{K}^{(j)}))$;
- **for** $t \leftarrow 1, 2, ..., T$ **do**
  - Tính S_W^A dựa trên (2.9) và $S_{W'}^A$ dựa trên (2.10) *theo thứ tự* O_A;
  - Giải bài toán tối ưu (2.8) bằng SDP để tìm β;
  - Tính S_W^β dựa trên (2.12) và $S_{W'}^\beta$ dựa trên (2.13) *theo thứ tự* O_β;
  - Giải bài toán giá trị riêng tổng quát (2.11) để tìm A;
- **return** A, β;

Độ phức tạp của Thuật toán 2 là $\mathcal{O}(N^3)$. Tuy nhiên, trong công thức trên N (số lượng ma trận thành phần trong chuỗi tính tích) có giá trị nhỏ là bằng 4, do đó, thời gian tính toán của Thuật toán 2 là không đáng kể.

2.4.2. Đề xuất Thuật toán fMKL-DR

Từ Thuật toán 2, ta tìm ra được thứ tự tích tích chuỗi ma trận để tối ưu hoạt hoạt động nhân. Từ đó, tác giả đề xuất thuật toán hiệu quả dựa trên giảm chiều dữ liệu kết hợp học đa nhân như Thuật toán 3. Thuật toán 3 được xây dựng dựa trên Thuật toán 1 với cùng chung bộ Input và Output. Trong đó, Thuật toán 3 đã được thực hiện tính toán

thứ tự nhân tốt nhất tại dòng 3, 4 nhằm tìm ra thứ tự nhân tích chuỗi ma trận O_A và O_β sao cho số phép nhân cần thực hiện là nhỏ nhất. Từ đó, các ma trận được tính tại dòng 6 và dòng 8 dựa trên thứ tự O_A và O_β. Các ma trận trên được tính lặp lại T lần, do đó, thời gian tính toán của Thuật toán 3 sẽ được giảm đi đáng kể.

2.5. Đề xuất mô hình phân lớp bệnh nhân dựa trên fMKL-DR

Hình 2.3: Mô hình hiệu quả phân lớp bệnh nhân dựa trên fMKL-DR

Tác giả đề xuất mô hình phân lớp bệnh nhân ung thư được thể hiện trong Hình 2.3. Mô hình tổng quát gồm năm bước sau:

1. **Chọn dữ liệu sinh học từ các nguồn khác nhau.**
2. **Tiền xử lý dữ liệu.**
3. **Tạo ma trận nhân cho mỗi loại dữ liệu.**
4. **Hợp nhất các loại dữ liệu.**
5. **Phân tầng bệnh nhân.**

Năm bước trên đây là mô hình tổng quát, chi tiết việc áp dụng các bước trong từng mô hình cụ thể đối với bệnh ung thư và bệnh Alzheimer được trình bày trong các thực nghiệm ở mục tiếp theo.

2.6. Thực nghiệm và kết quả

2.6.1. Tập dữ liệu

*** Tập dữ liệu bệnh nhân ung thư**

Với các thực nghiệm trên tập dữ liệu bệnh nhân ung thư, tác giả sử dụng các tập dữ liệu các bệnh ung thư khác nhau được tải về từ Thư viện bản đồ gen bệnh ung thư (The Cancer Genomie Atlas - TCGA[1] 2018). Đây là một bản đồ toàn diện, đa chiều về những thay đổi

[1]https://www.cancer.gov/tcga

di truyền quan trọng của 33 loại ung thư. Bộ dữ liệu từ TCGA gồm hơn 2 PetaByte dữ liệu di truyền được cung cấp công khai hỗ trợ cộng đồng nghiên cứu ung thư trên thế giới. Tác giả sử dụng 4 tập dữ liệu ung thư khác nhau từ TCGA, thuộc tính phân lớp cho mô hình là thuộc tính *Đã chết* (Dead). Chi tiết dữ liệu trong Bảng 2.1.

Tập dữ liệu	**Số lượng mẫu**	**Số đặc trưng trong mỗi loại dữ liệu**			
		Biểu hiện Gene	Methyl hóa DNA	Biểu hiện miRNA	Biểu hiện Protein
Ung thư phổi (LUNG)	106	12.042	23.074	352	218
Ung thư não (BGM)	275	12.042	22.896	534	218
Ung thư biểu mô vú (BREAST)	435	12.042	24.978		218
Ung thư buồng trứng (OV)	541	12.042	21.825	799	218

Bảng 2.1: Chi tiết tập dữ liệu bệnh nhân ung thư

*** Tập dữ liệu bệnh nhân Alzheimer**

Tác giả sử dụng tập dữ liệu ảnh cộng hưởng từ của bệnh nhân Alzheimer được tải về từ Alzheimer's Disease Neroimaging Initiative[2] (ADNI). ADNI là một sáng kiến nhằm cải thiện các thử nghiệm lâm sàng để chuẩn đoán và điều trị bệnh Alzheimer dựa trên hình ảnh. Kho dữ liệu của ADNI với dữ liệu đến từ khoảng 1.000 người tình nguyện bao gồm các nhóm trạng thái của bệnh Alzheimer (AD) như: khỏe mạnh, suy giảm nhận thức nhẹ, bị bệnh. Tập dữ liệu gồm 710 đối tượng quan sát ở mức trọng số T1 gồm: 200 đối tượng mắc AD; 280 đối tượng có biểu hiện suy giảm nhận thức nhẹ (Mild Cognitive Impairment - MCI), trong đó 120 đối tượng chuyển sang AD (MCI converted - MCIc) và 160 đối tượng không chuyển sang AD (MCI not converted - MCInc) sau 18 tháng; và 230 đối tượng khỏe mạnh (Normal Control - NC). Tất cả ảnh cộng hưởng từ của 710 đối tượng được tiền xử lý bằng phần mềm FreeSurfer v6.0. Chi tiết tập dữ liệu được thể hiện trong Bảng 2.2.

Loại	**Số lượng**	**Tuổi**	**Giới tính**	**MMSE**
NC	230	77,13±6,24	116/84	29,16±0,82
MCInc	160	76,26±5,35	89/71	27,56±1,18
MCIc	120	75,95±6,27	53/67	26,38±1,52
MCI	280	76,11±5,98	142/138	26,97±1,34
AD	200	76,63±5,91	122/108	23,54±2,07

Bảng 2.2: Chi tiết thông tin tập dữ liệu ảnh cộng hưởng từ bệnh nhân Alzheimer

2.6.2. Thực nghiệm 2.1: So sánh kết quả phân lớp từng loại dữ liệu riêng rẽ và tập dữ liệu tích hợp

Thực nghiệm này tác giả muốn đánh giá kết quả phân lớp và khi thực hiện phân lớp trên từng loại dữ liệu riêng lẻ với tập dữ liệu tích hợp cũng như sự đóng góp của từng loại dữ liệu vào tập dữ liệu tích hợp dựa trên tập dữ liệu bệnh nhân ung thư.

2.6.3. Thực nghiệm 2.2: Đánh giá hiệu năng thuật toán fMKL-DR với MKL-DR

Thực nghiệm này tác giả muốn đánh giá hiệu năng thực hiện của thuật toán fMKL-DR với các thuật toán MKL-DR và rMKL-DR về mặt thời gian thực hiện thuật toán. Dữ liệu

[2]http://adni.loni.usc.edu

của thực nghiệm này tác giả sử dụng tập dữ liệu bệnh nhân ung thư với 4 loại bệnh ung thư có số lượng mẫu quan sát từ thấp nhất với bệnh Ung thư phổi (106 mẫu) đến cao nhất là bệnh Ung thư buồng trứng (541 mẫu). Việc chọn các loại bệnh với số lượng mẫu quan sát từ thấp đến cao giúp đánh giá toàn diện và đầy đủ về thời gian thực hiện thuật toán trên các tập dữ liệu kích thước khác nhau. Mỗi loại bệnh ung thư, tác giả chọn 3 loại dữ liệu gồm: biểu hiện gene, methyl hóa DNA, biểu hiện miRNA.

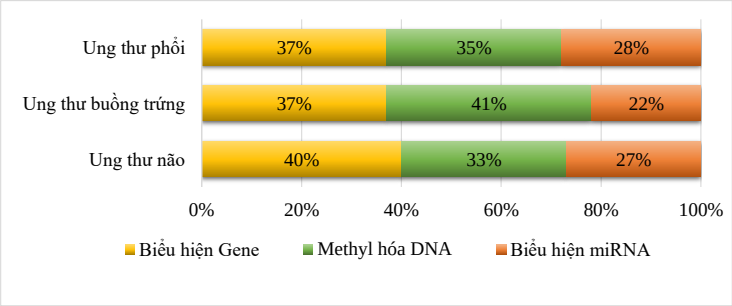
2.6.4. Thực nghiệm 2.3: Đánh giá hiệu năng của thuật toán fMKL-DR trên tập dữ liệu bệnh Alzheimer

Thực nghiệm này tác giả muốn đánh giá kết quả khi áp dụng fMKL-DR trên tập dữ liệu ảnh y tế, cụ thể là ảnh cộng hưởng từ của bệnh Alzheimer.

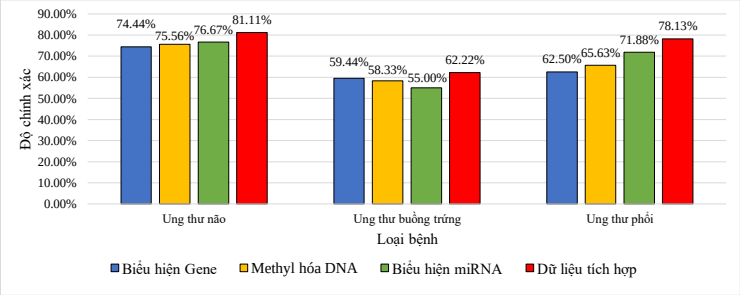
2.6.5. Kết quả và thảo luận

Qua thực hiện ba thực nghiệm, tác giả thu được một số kết quả và có một số đánh giá như sau:

Thực nghiệm thứ nhất: với mục tiêu là so sánh hiệu suất phân lớp khi phân lớp trên từng loại dữ liệu với tập dữ liệu thống nhất (được tích hợp dựa trên phương pháp MKL-DR) kết quả thể hiện hai điểm nổi bật:



Hình 2.5: Tỉ lệ đóng góp của các loại dữ liệu cho tập dữ liệu thống nhất

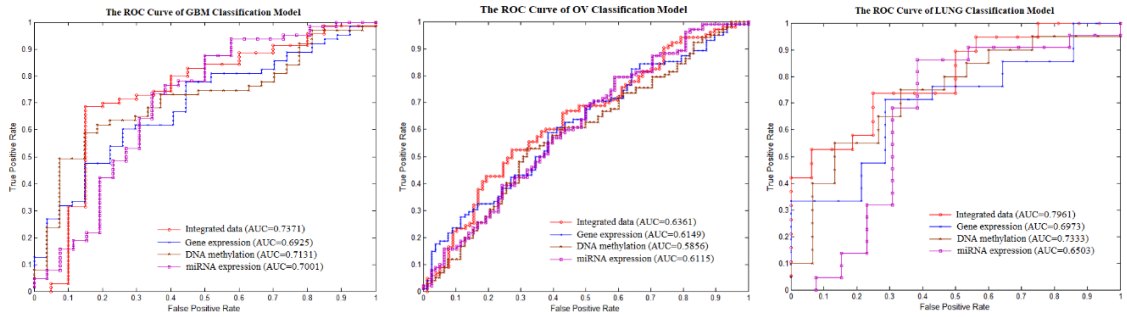


Hình 2.6: So sánh độ chính xác của bộ phân lớp trên từng tập dữ liệu

Thực nghiệm thứ hai: với mục tiêu so sánh thời gian thực hiện thuật toán giữa các phương pháp MKL-DR, rMKL-DR, fMKL-DR. Kết quả thể hiện trong Bảng 2.5.

Hình 2.8 thể hiện sự so sánh thời gian thực hiện giữa các phương pháp với cùng một số lần lặp là 10.

Thực nghiệm thứ ba, tác giả muốn chứng minh phương pháp fMKL-DR có thể áp dụng khả thi đối với bài toán xử lý ảnh y tế. Kết quả áp dụng thực nghiệm thứ ba trên



Hình 2.7: Đường cong ROC so sánh các mô hình phân lớp

Bệnh ung thư	**Thời gian (giây)**								
	5 lần lặp			**10 lần lặp**			**20 lần lặp**		
	MKL-DR	rMKL-DR	fMKL-DR	MKL-DR	rMKL-DR	fMKL-DR	MKL-DR	rMKL-DR	fMKL-DR
LUNG	4	4	**3**	9	9	**6**	19	19	**13**
GBM	69	69	**41**	138	139	**82**	276	279	**165**
BREAST	1.064	1.064	**714**	2.130	2.203	**1.428**	4.262	4.409	**2.857**
OV	1.716	1.750	**1.237**	3.433	3.502	**2.475**	6.867	7.005	**4.952**

Bảng 2.5: Kết quả thời gian thực hiện các phương pháp trong thực nghiệm 2.2

tập dữ liệu ảnh cộng hưởng từ của bệnh nhân Alzheimer đã cho kết quả tốt. Kết quả thể hiện trong Bảng 2.6.

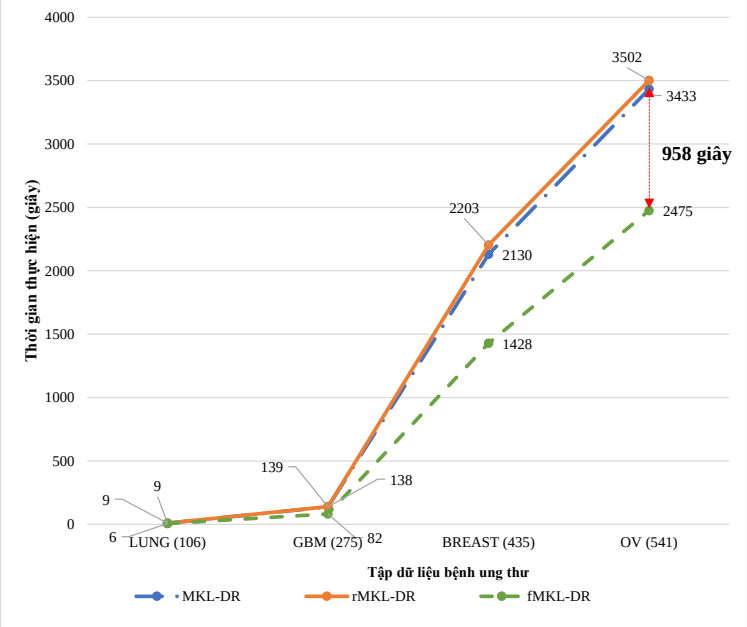
Phương pháp	**AD/NC**		**AD/MCI**		**NC/MCI**		**MCIc/MCInc**	
	ACC (%)	AUC	ACC (%)	AUC	ACC (%)	AUC	ACC (%)	AUC
Chupin, 2009	80,51	0.7851	73,48	0,7328	71,94	0,7155	64,21	0,6638
Alhed, 2015	86,40	0,8487	74,51	0,7562	76,29	0,7677	68,72	0,6814
Khedher, 2015	88,96	0,9256	84,59	0,8859	82,41	0,8134	70,11	0,7076
Dai, 2013	90,81	0,9429	85,92	0,8743	81,92	0,8118	71,04	0,7086
Suk, 2014	93,05	0,9475	88,98	0,9007	83,67	0,8203	72,86	0,7123
Liu, 2017	95,24	0,9754	90,85	0,9355	86,35	0,9107	74,28	0,7885
fMKL-DR	**96,28**	**0,9782**	**91,04**	**0,9413**	**87,84**	**0,9148**	**78,21**	**0,8019**

Bảng 2.6: Kết quả phân lớp trên tập dữ liệu bệnh Alzheimer

2.7. Kết luận

Trong chương này, luận án đã trình bày những nội dung, kết quả nghiên cứu về phương pháp Giảm chiều dữ liệu kết hợp học đa nhân. Ngoài ra tác giả đã đề xuất một phương pháp để tăng hiệu năng tính toán và đặt tên là Thuật toán hiệu quả dựa trên giảm chiều dữ liệu kết hợp học đa nhân. Bằng phương pháp phân tích, đề xuất giải pháp và thực nghiệm, luận án đã thể hiện những kết quả sau:

Thứ nhất, chứng minh được MKL-DR không chỉ áp dụng hiệu quả trên tập dữ liệu hình ảnh mà còn phù hợp với tập dữ liệu Tin-sinh học, cụ thể là tập dữ liệu bệnh nhân ung thư. Kết quả thực nghiệm đã chỉ ra, tích hợp dữ liệu kết hợp với giảm chiều dữ liệu phù hợp với tập dữ liệu bệnh nhân ung thư do tập dữ liệu này có nhiều loại dữ liệu được thu thập, mỗi loại dữ liệu lại có số chiều lớn nên MKL-DR đã tỏ rõ tác dụng.



Hình 2.8: So sánh thời gian thực hiện khi số lần lặp là 10 với kích thước tập dữ liệu khác nhau

Thứ hai, đề xuất được một thuật toán (fMKL-DR) giúp giảm đáng kể thời gian tính toán của phương pháp mà kết quả mang lại vẫn tương đương. Phương pháp này có ý nghĩa rất lớn khi thực tế hiện nay có rất nhiều dữ liệu được quan sát, số mẫu ngày càng lớn nên hiệu quả về mặt thời gian cũng là một yếu tố cần phải xem xét khi đánh giá một thuật toán.

Thứ ba, đề xuất xây dựng mô hình phân lớp bệnh nhân Alzheimer dựa trên ảnh cộng hưởng từ. Mô hình được đề xuất đã cho kết quả tốt khi thực nghiệm trên một tập tương đối lớn ảnh cộng hưởng từ của bệnh Alzheimer. Mô hình này sẽ trợ giúp cho việc chuẩn đoán sớm bệnh Alzheimer - một trong những yếu tốt then chốt trong việc điều trị bệnh Alzheimer.

Chương 3

PHÂN LỚP BỆNH NHÂN DỰA TRÊN PHƯƠNG PHÁP PHÂN TÍCH THÀNH PHẦN CHÍNH TĂNG CƯỜNG

Chương này, tác giả tập trung trình bày phương pháp RPCA và đề xuất một cách tiếp cận mới trong việc ứng dụng RPCA vào xử lý tập dữ liệu sinh học phân tử. Từ đó, tác giả xây dựng mô hình phân lớp bệnh nhân ung thư dựa trên RPCA và thu được một bộ công cụ phân lớp bệnh nhân ung thư, trợ giúp trong quá trình điều trị bệnh nhân ung thư. Các kết quả của Chương này được công bố trên bài báo [GTTrung-5].

3.1. Giới thiệu

3.2. Phương pháp phân tích thành phần chính

3.2.1. Giới thiệu

3.2.2. Phương pháp PCA

3.3. Phương pháp phân tích thành phần chính tăng cường

3.3.1. Giới thiệu

Nhược điểm của phương pháp PCA là rất nhạy cảm với lỗi lớn (hay còn gọi là dữ liệu ngoại lai - các phần tử bị lỗi, có giá trị nằm quá xa trung bình dữ liệu, các lỗi này có thể do bị tác động từ con người hoặc lỗi khi đo đạc). Candes và cộng sự trong đề xuất một phương pháp cải tiến phương pháp phân tích thành phần chính để thích nghi với dữ liệu ngoại lai gọi là Phương pháp phân tích thành phần chính tăng cường (Robust Principal Component Analysis - RPCA). RPCA đã được áp dụng trên nhiều bài toán như: thị giác máy, gióng ảnh, khôi phục không gian con, phân cụm. PCA xuất phát từ ý tưởng phân rã ma trận dữ liệu ban đầu $O \in \mathbb{R}^{m\times n}$ thành tổng của hai ma trận $O = L + N$, trong đó có một ma trận hạng thấp L (chứa phần lớn thông tin gọi là các thành phần chính) và một ma trận nhiễu N. RPCA cũng xuất phát từ ý tưởng đó khi phân rã $O = L + S$, tuy nhiên, trong khi PCA xác định N nhỏ để giảm thiểu mất mát thông tin thì S là ma trận thưa có thể có độ lớn tùy ý. Bài toán RPCA trở thành bài toán xác định ma trận hạng thấp L từ phân rã $O = L + N$.

Bài toán RPCA được đưa về bài toán tối ưu dựa trên ℓ_0-norm sau:

$$\begin{aligned} \underset{L,S}{\text{minimize}} \quad & rank(L) + \lambda||S||_0 \\ \text{subject to} \quad & O = L + S \end{aligned} \tag{3.14}$$

với $\lambda > 0$ là nhân tử Lagrange. Tuy nhiên, bài toán (3.14) là bài toán NP-khó.

3.3.2. Các hướng giải bài toán RPCA

Đã có nhiều phương pháp được đề xuất nhằm giải bài toán (3.14), bài toán phân rã ma trận (3.14) được biến đổi dựa trên phương pháp theo đuổi thành phần chính (Principal Component Persuit - PCP), sau đó bài toán được giải dựa trên các phương pháp Gia tăng độ dốc gần (Accelerated Proximal Gradient - APG) hay Nhân tử Lagrange tăng cường (Augmented Lagrange Multipliers).

Bài toán phân rã ma trận RPCA (3.14) được giải thông qua bài toán Theo đuổi thành phần chính (PCP) sau:

$$\begin{aligned} \underset{L,S}{\text{minimize}} \quad & ||L||_* + \lambda||S||_1 \\ \text{subject to} \quad & O = L + S \end{aligned} \tag{3.15}$$

với $||X||_* = \sum_i \sigma_i(X)$ là nuclear norm của ma trận X (là tổng các giá trị căn bậc hai của các trị riêng (singular value) của X); $||X||_1 = \sum_{ij} |X_{ij}|$ là ℓ_1-norm của X (được xem như là một vector dài trong $\mathbb{R}^{m \times n}$); $\lambda = 1/\sqrt{\max(m,n)}$. Bài toán (3.15) có thể được biến đổi về dạng bài toán Semidefinite Program (SDP) và giải qua phương pháp điểm phía trong.

3.3.3. Phương pháp dựa trên Nhân tử Lagrange tăng cường

3.3.3.1. Phương pháp Nhân tử Lagrange tăng cường

3.3.3.2. Thuật toán Exact Augmented Lagrange Multiplier

3.3.3.3. Thuật toán Inexact Augmented Lagrange Multiplier (IALM)

3.3.3.4. Một số lưu ý khi áp dụng thuật toán EALM và IALM

3.3.4. Đánh giá phương pháp RPCA

a, Ưu điểm

Phương pháp RPCA có một số ưu điểm sau:

- Mạnh mẽ với các tập dữ liệu quan sát có nhiều dữ liệu ngoại lai. RPCA có thể phân tách riêng các dữ liệu thưa ra khỏi tập dữ liệu chính xác giúp gia tăng hiệu quả của các phương pháp phân tích dữ liệu.
- Các mẫu thưa thớt của ma trận thưa S không cần xác định trước, phụ thuộc vào dữ liệu.
- Có thể mở rộng một cách tự nhiên cho bài toán hoàn thành ma trận (Matrix Completion).
- Ứng dụng vào các bài toán khác ngoài bài toán thị giác máy. Tuy nhiên, cách nhìn mới không còn là phân rã tìm ra ma trận nhiễu hoặc lỗi mà có thể nhìn theo cách là tìm ra những điểm khác biệt so với phần còn lại. Có thể xem những thành phần khác 0 trong ma trận thưa chính là phần dữ liệu khác biệt, khi đó bài toán phân tích dữ liệu có thể chỉ cần xem xét trên phần dữ liệu khác biệt đó.

b, Hạn chế

Ngoài những ưu điểm, RPCA cũng thể hiện một số hạn chế sau:

- Tốc độ hội tụ của phương pháp là nhược điểm lớn nhất của RPCA. Do thường xuyên phải lặp đi lặp lại trong việc tính norm của các ma trận trong khi giải các bài toán tối ưu nên làm giảm hiệu năng về mặt thời gian tính toán. Một số phương pháp đề xuất giảm bớt việc tính toán norm của các ma trận khi giảm số vòng lặp để tối ưu hóa nhưng cũng làm ảnh hưởng đến độ chính xác của các mô hình phân tích dữ liệu sau này. Do gặp vấn

đề về mặt hiệu năng tính toán nên đối với các bài toán thị giác máy, RPCA chưa thực sự hiệu quả khi các hình ảnh hoặc video có độ phân giải cao.

- Tiềm năng của RPCA là khá lớn khi ngày càng được ứng dụng trong nhiều lĩnh vực (bài toán) khác nhau nhưng chưa nhiều. Do vậy, các mô hình bài toán mới dựa trên RPCA cần được nghiên cứu để ứng dụng RPCA trong nhiều lĩnh vực khác ngoài một số lĩnh vực mà RPCA được phát triển dựa trên đó.

3.3.5. Các hướng nghiên cứu mở rộng và áp dụng RPCA

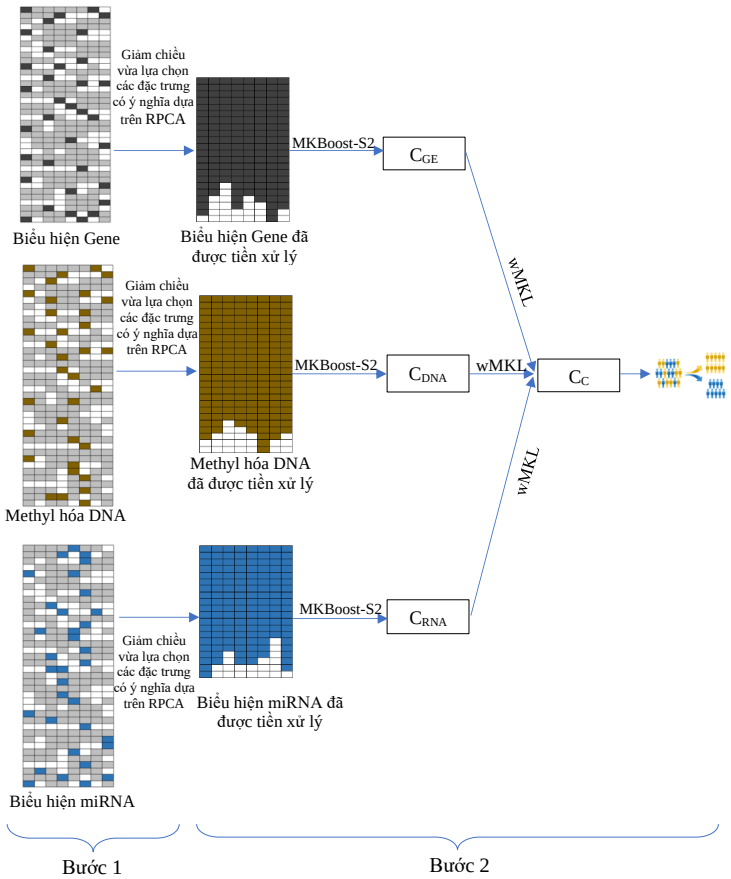
3.3.5.1. Cải thiện hiệu năng thuật toán

3.3.5.2. Áp dụng RPCA vào các bài toán ứng dụng Tin-sinh học

3.3.5.3. Áp dụng RPCA vào các bài toán phân tích hình ảnh y sinh

3.4. Đề xuất mô hình phân lớp bệnh nhân dựa trên phương pháp phân tích thành phần chính tăng cường

Tác giả đề xuất mô hình phân lớp bệnh nhân ung thư kết hợp mô hình giảm chiều dữ liệu để lựa chọn các đặc trưng có ý nghĩa với xây dựng mô hình phân lớp dựa trên học đa nhân theo mô hình trong Hình 3.5 dưới đây:



Hình 3.5: Mô hình phân lớp bệnh nhân ung thư dựa trên RPCA

Mô hình phân lớp bệnh nhân ung thư được đề xuất gồm 2 bước: Bước 1 thực hiện giảm chiều dữ liệu và trích chọn các đặc trưng có ý nghĩa dựa trên RPCA từ các tập dữ liệu gốc (chi tiết được trình bày trong mục 3.4.1); Bước 2 xây dựng các bộ phân lớp dựa trên

MKBoost-S2 và tích hợp các bộ phân lớp bằng wMKL để tạo thành bộ phân lớp bệnh nhân ung thư (chi tiết được trình bày trong mục 3.4.2). Dữ liệu đầu vào sẽ là các tập dữ liệu sinh học phân tử (biểu hiện gene, methyl hóa DNA, biểu hiện miRNA) của các bệnh nhân ung thư, dữ liệu đầu ra của mô hình sẽ là các bệnh nhân ung thư được phân vào các lớp khác nhau.

3.4.1. Giảm chiều dữ liệu kết hợp chọn lọc các đặc trưng khác biệt dựa trên RPCA

Tác giả đề xuất mô hình phân rã dữ liệu dựa trên RPCA đối với dữ liệu biểu hiện gene (tương tự với hai loại dữ liệu methyl hóa DNA và biểu hiện miRNA) như sau:
- **Bước 1.** Phân rã ma trận dữ liệu gốc
- **Bước 2.** Sắp xếp các gen dựa trên giá trị
- **Bước 3.** Trích xuất các gene có ý nghĩa

3.4.2. Phân lớp dựa trên Học đa nhân

Từ đó, tác giả đề xuất mô hình phân lớp bệnh ung thư từ nhiều nguồn dữ liệu khác nhau gồm 2 bước MKL thể hiện trong Bước 2 của Hình 3.5.

3.5. Thực nghiệm và kết quả

3.5.1. Tập dữ liệu

Tác giả sử dụng các tập dữ liệu các bệnh ung thư khác nhau được tải về từ Thư viện bản đồ gen bệnh ung thư (The Cancer Genomie Atlas - TCGA[1] 2018) gồm 4 tập dữ liệu ung thư khác nhau từ TCGA, thuộc tính phân lớp cho mô hình là thuộc tính Đã chết (Dead). Các tập dữ liệu bệnh nhân ung thư bao gồm: Lung Squamous Cell Carcinoma (LUNG), Glioblastoma Multiforme (GBM), Breast Invasive Carcinoma (BREAST), Ovarian Serous Cytadenocarcinoma (OV).

Với mỗi tập dữ liệu bệnh nhân ung thư, sử dụng ba loại dữ liệu liên quan gồm: dữ liệu biểu hiện gene, methyl hóa DNA và dữ liệu biểu hiện miRNA. Chi tiết dữ liệu được thể hiện trong Bảng 3.1.

Tập dữ liệu	**Số lượng mẫu**	**Còn sống / Đã chết**	**Số đặc trưng trong mỗi loại dữ liệu**		
			Biểu hiện Gene	Methyl hóa DNA	Biểu hiện miRNA
Ung thư phổi (LUNG)	106	42/64	12.042	23.074	352
Ung thư não (BGM)	275	202/73	12.042	22.896	534
Ung thư vú (BREAST)	435	360/75	12.042	21.825	799
Ung thư buồng trứng (OV)	541	283/258	12.042	21.825	799

Bảng 3.1: Tập dữ liệu bệnh nhân ung thư

3.5.2. Thiết kế thực nghiệm

Tác giả thiết kế thực nghiệm để đánh giá hiệu quả của phương pháp theo các bước dưới đây:

[1] https://www.cancer.gov/tcga

Đầu tiên, mỗi tập dữ liệu trong các tập biểu hiện gene, methyl hoá DNA, biểu hiện miRNA được biểu diễn như một ma trận O. Mỗi dòng tương ứng với một đặc trưng, mỗi cột là một mẫu quan sát tương ứng với một bệnh nhân. Sau đó, tác giả áp dụng mô hình giảm chiều kết hợp lựa chọn đặc trưng khác biệt dựa trên RPCA được đề xuất ở Mục 3.4.1 vào từng tập dữ liệu. Kết quả thu được 3 ma trận dữ liệu đã được tiền xử lý.

Tiếp theo, áp dụng MKBoost-S2 tạo các bộ phân lớp C_{GE}, C_{DNA}, C_{RNA} từ các tập dữ liệu tương ứng. Từ mỗi tập dữ liệu, sử dụng 13 hàm nhân cơ sở nhằm tăng hiệu năng cho phương pháp MKBoost-S2. Sau đó, sử dụng wMKL để kết hợp 3 bộ phân lớp CGE, CDNA, CRNA thành một bộ phân lớp tổng hợp duy nhất C_C. Để đánh giá hiệu quả của việc tích hợp, tác giả còn thực hiện kết hợp từng cặp 2 bộ phân lớp để tạo ra các bộ phân lớp $C_{GE\text{-}DNA}$, $C_{GE\text{-}RNA}, C_{DNA\text{-}RNA}$.

Tác giả đánh giá kết quả dựa trên độ chính xác và đường cong ROC. Tác giả thực hiện 2 so sánh để đánh giá như sau: Thứ nhất, so sánh kết quả phân lớp trên tập dữ liệu đã được tiền xử lý bằng RPCA với tập dữ liệu gốc; Thứ hai, so sánh các bộ phân lớp khi áp dụng mô hình đề xuất khi chỉ tích hợp 2 trong 3 bộ phân lớp so với tích hợp cả 3 bộ phân lớp. Với mỗi mô hình phân lớp, thực hiện chạy 20 lần Bước thứ hai để xây dựng bộ phân lớp và xác định độ chính xác của các mô hình. Mỗi lần chạy, lấy ngẫu nhiên 2/3 tập dữ liệu để huấn luyện mô hình và 1/3 tập dữ liệu còn lại để kiểm thử mô hình. Độ chính xác của từng mô hình phân lớp được tính bằng trung bình cộng 20 lần chạy.

3.5.3. Kết quả thực nghiệm

Bệnh	**Số lượng mẫu**	**Độ chính xác (%)**					
		Dữ liệu gốc			Dữ liệu được xử lý bởi RPCA		
		Biểu hiện Gene	Methyl hóa DNA	Biểu hiện miRNA	C_{GE}	C_{DNA}	C_{RNA}
LUNG	106	61.88	64.85	70.94	**64.68**	**67.81**	**71.41**
BGM	275	74.22	75.28	76.39	**80.83**	**76.39**	**80.88**
BEAST	435	88.10	88.03	91.48	**90.14**	**90.10**	**91.51**
OV	541	59.22	58.22	54.92	**68.72**	**67.61**	**66.50**

Bảng 3.2: Độ chính xác của các bộ phân lớp giữa các tập dữ liệu gốc và các tập dữ liệu được tiền xử lý dựa trên RPCA

Bệnh	**Số lượng mẫu**	**Độ chính xác (%)**				
		$C_{GE\text{-}DNA}$	$C_{GE\text{-}RNA}$	$C_{DNA\text{-}RNA}$	C_C	C_C^*
LUNG	106	69.22	72.66	72.35	**77.35**	*76.65*
BGM	275	81.67	82.72	81.61	**85.23**	*84.80*
BEAST	435	91.44	91.43	92.17	**92.92**	*92.73*
OV	541	69.25	68.70	67.64	**69.80**	*69.56*

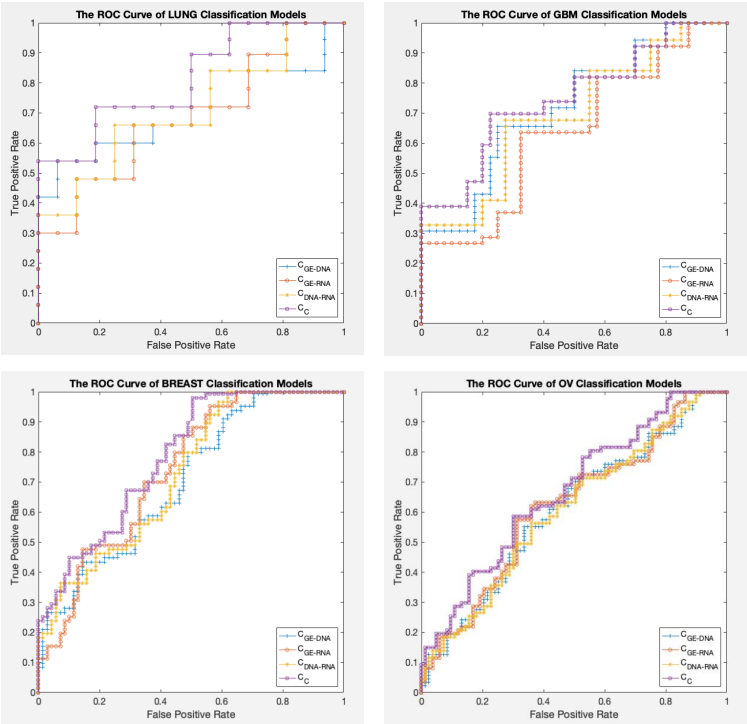
Bảng 3.3: Độ chính xác của các mô hình phân lớp tích hợp từ 2, 3 bộ phân lớp thành phần

Độ chính xác của các bộ phân lớp được thể hiện trong Bảng 3.2 và Bảng 3.3.

Bảng 3.4 thể hiện các giá trị AUC của các mô hình phân lớp và Hình 3.7 biểu diễn Biểu đồ đường cong ROC của chúng.

3.6. Kết luận

Trong chương này, tác giả đã trình bày những nội dung, kết quả nghiên cứu về phương pháp Phân tích thành phần chính tăng cường (RPCA). Ngoài ra tác giả đã đề xuất một



Hình 3.7: Biểu đồ đường cong ROC của các mô hình phân lớp trên từng tập dữ liệu bệnh ung thư

Bệnh	**AUC**			
	$C_{GE\text{-}DNA}$	$C_{GE\text{-}RNA}$	$C_{DNA\text{-}RNA}$	C_C
LUNG	0.7324	0.7093	07225	**0.8135**
BGM	0.7383	0.7066	0.7251	**0.7683**
BEAST	0.7241	0.7624	0.7498	**0.7925**
OV	0.6217	0.6255	0.6132	**0.6746**

Bảng 3.4: Giá trị AUC của các mô hình phân lớp

mô hình phân lớp cho một số bài toán Tin-sinh học.

Bằng phương pháp phân tích, đề xuất giải pháp và thực nghiệm, chương này của luận án đã thể hiện những kết quả sau:

Thứ nhất, trình bày một cách chi tiết phương pháp Phân tích thành phần chính, những ưu điểm, nhược điểm của phương pháp. Sau đó là phương pháp mở rộng phương pháp Phân tích thành phần chính gọi là Phân tích thành phần chính tăng cường (RPCA) từ ý tưởng, các hướng giải bài toán và 2 thuật toán hiệu quả là EALM và IALM để giải quyết bài toán. Ngoài ra, luận án trình bày các hướng mở rộng và khả năng ứng dụng của phương pháp RPCA trong tương lai. Bao gồm 3 hướng chính là: cải thiện hiệu năng thuật toán, kiểm tra tính ứng dụng trong các bài toán Tin-sinh học và các bài toán phân tích ảnh y sinh.

Thứ hai, đề xuất thực nghiệm xây dựng mô hình phân lớp bệnh nhân ung thư dựa trên tập dữ liệu biểu hiện gen. Ứng dụng của phương pháp RPCA trong bài toán này là việc phân tách tập dữ liệu biểu hiện gen ban đầu thành các gen có sự khác biệt trong tập dữ liệu, từ đó sàng lọc và phân tích dựa trên dữ liệu khác biệt và có ý nghĩa phân tách hơn so với tập dữ liệu gốc. Kết quả thực nghiệm cho thấy, mô hình đề xuất cho kết quả phân lớp tốt hơn so với tập dữ liệu gốc.

KẾT LUẬN

Các kết quả đạt được

Giảm chiều dữ liệu đã trở thành một bước tiền xử lý đóng vai trò hết sức quan trọng trong quá trình Khai phá tri thức từ dữ liệu ở nhiều lĩnh vực. Đã có nhiều phương pháp giảm chiều dữ liệu được đề xuất nhằm biến đổi dữ liệu từ không gian có số chiều cao (với nhiều tồn tại như không phù hợp với mô hình tính toán, chứa nhiều nhiễu, dữ liệu thưa) sang không gian có số chiều thấp hơn (phù hợp với mô hình tính toán, loại bỏ nhiễu, cô đặc dữ liệu). Tuy nhiên, các nghiên cứu cũng phải đối mặt với nhiều thách thức mới khi tập dữ liệu ngày càng trở nên "lớn", đa dạng về loại hình và phức tạp trong các mối quan hệ. Để thích nghi được với những thách thức mới đặt ra, các phương pháp đã đề xuất cần được cải tiến để phù hợp với sự "lớn" và đa dạng của dữ liệu. Trong phạm vi nghiên cứu của mình, tác giả đặt ba mục tiêu chính và kết quả đạt được của luận án như sau:

- Thứ nhất, luận án đã trình bày được tổng quan về các phương pháp giảm chiều dữ liệu, các phương pháp nổi bật đã được đề xuất và ứng dụng trong xử lý các tập dữ liệu Tin-sinh học. Các phương pháp được phân loại vào các nhóm với chiến lược tiếp cận khác nhau, luận án cũng đã làm nổi bật được những ưu, nhược điểm của từng nhóm phương pháp cũng như thảo luận về hướng mở rộng các nhóm phương pháp đã được đề xuất.
- Thứ hai, luận án đã nghiên cứu chi tiết hai phương pháp giảm số chiều hiệu quả là MKL-DR và RPCA, phân tích những ưu, nhược điểm của hai phương pháp. Nghiên cứu cải tiến MKL-DR bằng cách cải thiện hiệu năng về thời gian tính toán, từ đó đề xuất phương pháp fMKL-DR. Phương pháp được đề xuất đã cải thiện đáng kể thời gian thực hiện phương pháp, fMKL-DR đã tăng tính ứng dụng của phương pháp khi các tập dữ liệu hiện nay ngày càng có xu hướng ngày càng tăng về số lượng mẫu quan sát cũng như số lượng đặc trưng.
- Thứ ba, luận án đã xây dựng được hai mô hình nhằm phân lớp bệnh nhân. Các mô hình này áp dụng các phương pháp fMKL-DR và RPCA nhằm giảm chiều dữ liệu. Không chỉ dừng ở giảm chiều dữ liệu, mô hình phân lớp được đề xuất còn bao gồm cả việc tích hợp dữ liệu từ nhiều nguồn khác nhau nhằm tận dụng thông tin hữu ích trong từng loại dữ liệu riêng rẽ. Mô hình phân lớp bệnh nhân được đề xuất rất phù hợp khi hiện nay, mỗi đối tượng thường được quan sát ở nhiều khía cạnh, mỗi khía cạnh lại mang những thông tin hữu ích khác nhau, do đó, tích hợp thông tin từ các tập dữ liệu khác nhau đang là xu hướng của các phương pháp phân tích dữ liệu hiện đại.

Hạn chế và hướng nghiên cứu tiếp theo

Hạn chế

Số bài toán đặt ra và lĩnh vực ứng dụng của các phương pháp giảm chiều dữ liệu trong thời đại của "dữ liệu lớn" là rất lớn, cần nhiều hơn nữa sự quan tâm nghiên cứu. Trong

nghiên cứu của mình, mặc dù đã đạt được một số kết quả nhất định, tuy nhiên, luận án cũng còn một số hạn chế sau:

- Thứ nhất, mặc dù các phương pháp được đề xuất để giảm chiều dữ liệu đã có hiệu quả nhất định trong việc giảm chiều các tập dữ liệu Tin-sinh học, kết quả phân lớp đã tốt hơn so với các tập dữ liệu không được áp dụng. Tuy nhiên, đối với một số tập dữ liệu bệnh nhân ung thư, kết quả phân lớp mới chỉ dừng lại ở mức chấp nhận được như đối với tập dữ liệu bệnh nhân ung thư buồng trứng.
- Thứ hai, hiện nay đã có thêm một số loại dữ liệu khác ngoài bốn loại dữ liệu biểu hiện gene, methyl hóa DNA, biểu hiện miRNA và biểu hiện Protein. Tuy nhiên, trong các thực nghiệm của luận án, tác giả chưa có điều kiện đưa thêm các loại dữ liệu khác vào để góp phần tăng độ chính xác của các mô hình đề xuất.

Hướng nghiên cứu tiếp theo

Trong thời gian tới, tác giả đề ra một số hướng nghiên cứu tiếp theo của mình liên quan đến lĩnh vực giảm chiều dữ liệu như sau:

- Tiếp tục nghiên cứu các phương pháp giảm chiều dữ liệu mới được đề xuất làm cơ sở để đề xuất phương pháp giảm chiều dữ liệu mới cũng như cải tiến các phương pháp giảm chiều dữ liệu đã có, tập trung vào các phương pháp giảm chiều dữ liệu ứng dụng được trong các bài toán Tin-sinh học.
- Xây dựng một số bộ công cụ giảm chiều dữ liệu phù hợp để xử lý các tập dữ liệu Sinh học phân tử. Trong đó, cơ sở lý thuyết sẽ dựa vào các phương pháp giảm chiều dữ liệu kết hợp tích hợp dữ liệu từ nhiều loại dữ liệu khác nhau để tạo nên tập dữ liệu thống nhất. Đặc biệt, việc kết hợp giữa dữ liệu ảnh và dữ liệu microarray trong chuẩn đoán, điều trị các bệnh là một hướng nghiên cứu tiềm năng do bản chất từng loại dữ liệu đó đã chứa rất nhiều thông tin hữu ích, nếu kết hợp chúng lại sẽ có khả năng nâng cao chất lượng của các mô hình phân tích dữ liệu.

Danh mục các công trình khoa học của tác giả liên quan đến luận án

GTTrung-1 . **Thanh Trung Giang**, Thanh Phuong Nguyen, and Dang Hung Tran (2017). "Stratifying cancer patients based on multiple kernel learning and dimensionality reduction", In 2017 9th International Conference on Knowledge and Systems Engineering (KSE), IEEE, pp. 106-111, IEEE Xplore. (Scopus, DBLP)

GTTrung-2 . **Thanh Trung Giang**, Thanh Phuong Nguyen, Quoc Vinh Nguyen Tran, and Dang Hung Tran (2018). "fMKL-DR: A Fast Multiple Kernel Learning Framework with Dimensionality Reduction", In International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making, Springer, Cham, pp. 153-165. (DBLP)

GTTrung-3 . **Thanh Trung Giang**, Thanh Phuong Nguyen, and Dang Hung Tran (2020). "Stratifying Patients Using Fast Multiple Kernel Learning Framework: Case studies of Alzheimer's Disease and Cancers", BMC Medical Informatics and Decision Making, 108 (2020). (SCIE Q1, IF = 2.067)

GTTrung-4 . **Thanh Trung Giang**, Thanh Phuong Nguyen, Quang Trung Pham, and Dang Hung Tran (2021). "A Combination Model of Robust Principal Component Analysis and Multiple Kernel Learning for Cancer Patient Stratification", The Second International Conference on Artificial Intelligence and Computational Intelligence (AICI 2021), Springer.

Danh mục này gồm 04 công trình.