Tích hợp dữ liệu sử dụng học đa nhân kết hợp với giảm chiều dữ liệu thực hiện phân cụm các bệnh nhân ung thư

Giang Thành Trung – Trường Đại học Tây Bắc, Trần Đăng Hưng – Trường Đại học Sư phạm Hà Nội

Tóm tắt

Mặc dù các nghiên cứu về ung thư đang diễn ra nhưng các phương pháp điều trị hiện có là rất hạn chế về số lượng và hiệu quả, ngoài ra quyết định điều trị cho các bệnh nhân vẫn là một vấn đề khó khăn. Việc thành lập các phân nhóm giúp trợ giúp ra quyết định chủ yếu vẫn dựa trên các loại dữ liệu riêng biệt. Tuy nhiên, việc phân tích dữ liệu bệnh nhân đa chiều dựa trên nhiều số đo đặc trưng phân tử khác nhau như: biểu thức Gene, DNA Methylation, biểu thức miRNA có thể giúp phát hiện các đặc điểm nội tại của các khối u chính xác hơn. Một phương pháp giúp tích hợp dữ liệu từ nhiều dạng đặc tả khác nhau dựa trên phương pháp học đa nhân đồng thời kết hợp với việc giảm chiều dữ liệu được chúng tôi sử dụng trong báo cáo này nhằm tạo nên một bộ dữ liệu có tính tổng quát cao hơn từ nhiều loại biểu diễn dữ liệu khác nhau. Sau khi tích hợp dữ liệu, chúng tôi tiến hành phân cụm bệnh nhân ung thư dựa trên dữ liệu đã tích hợp và đánh giá kết quả phân cụm thu được. Ngoài ra chúng tôi cũng đưa ra các gợi ý về mặt lựa chọn hàm nhân cho các loại dữ liệu cũng như các tham số để đạt được kết quả phân cụm tốt nhất cho mô hình này.

Mục tiêu

Xây dựng được mô hình tích hợp dữ liệu kết hợp với một phương pháp giảm chiều dữ liệu từ các loại dữ liệu biểu diễn các số đo đặc trưng phân tử khác nhau của bệnh nhân ung thư như: biểu thức Gene, DNA methylation, biểu thức miRNA.

Dựa trên dữ liệu đã tích hợp được để tiến hành phân cụm và đánh giá kết quả phân cụm, từ đó có những đề xuất về việc lựa chọn hàm nhân và các tham số cho hàm nhân cũng như cho mô hình huấn luyện.

Phương pháp

Để tích hợp dữ liệu từ các loại dữ liệu khác nhau chúng tôi sử dụng phương pháp "Học đa nhân kết hợp giảm chiều dữ liệu - Multiple Kernel Learning for Dimensionality Reduction (MKL-DR)" (Lin và các đồng sự, 2011). Phương pháp này một mặt tiến hành học đa nhân, mặt khác sử dụng mô hình "nhúng đồ thị" để giảm chiều dữ liệu. Phương pháp đó cụ thể như sau:

*** Học đa nhân**

Học đa nhân là việc tối ưu hóa các trọng số β mà nó là tổ hợp tuyến tính một tập các ma trận nhân $\{K_1, K_2, ..., K_M\}$, với M là số ma trận nhân tương ứng với số loại đặc tả để tạo nên một ma trận nhân thống nhất K sao cho:

$$K = \sum_{m=1}^{M} \beta_m K_m \quad (1)$$

Với mỗi loại dữ liệu được biểu diễn thành một ma trận nhân tương ứng.

*** Nhúng đồ thị**

MKL-DR sử dụng nhúng đồ thị như một nền tảng để giảm chiều dữ liệu (Yan và các đồng sự, 2007), nó cho phép áp dụng cho hầu hết các phương pháp giảm chiều phổ biến.

Trong nền tảng này, một vector chiếu v (để chiếu vào không gian 1 chiều) hay một ma trận chiếu V (để chiếu vào không gian nhiều chiều) được tối ưu dựa trên tiêu chuẩn bảo toàn đồ thị sau:

$$\underset{v}{\text{minimize}} \sum_{i,j=1}^{N} \left\| v^T x_i - v^T x_j \right\|^2 w_{ij} \quad (2)$$

$$s.t. \quad \sum_{i=1}^{N} \left\| v^T x_i \right\| d_{ii} = \text{const}$$

$$\text{or} \quad \sum_{i,j=1}^{N} \left\| v^T x_i - v^T x_j \right\|^2 w'_{ij} = \text{const}$$

với N là số mẫu dữ liệu, v là vector chiếu, W là ma trận tương tự với các phần tử w_{ij} , D (hoặc W') là ma trận ràng buộc nhằm tránh nghiệm tầm thường của hàm mục tiêu. Việc chọn W và D phụ thuộc vào phương pháp giảm chiều được sử dụng.

*** Học đa nhân kết hợp giảm chiều dữ liệu**

Khi sử dụng nhúng đồ thị với các phương pháp nhân ta có $v = \sum_{n=1}^{N} a_n \emptyset(x_n)$ thì bài toán (2) trở thành:

$$\underset{v}{\text{minimize}} \sum_{i,j=1}^{N} \left\| \alpha^T \mathcal{K}^i \beta - \alpha^T \mathcal{K}^j \beta \right\|^2 w_{ij} \quad (3)$$

$$s.t. \quad \sum_{i=1}^{N} \left\| \alpha^T \mathcal{K}^i \beta \right\|^2 d_{ii} = \text{const}$$

$$\beta_m \geq 0, m = 1, 2, ...M$$

với: $\alpha = [\alpha_1, \alpha_2, ...\alpha_N]^T \in \mathbb{R}^N, \beta = [\beta_1, \beta_2, ..., \beta_M]^T \in \mathbb{R}^M$

$$\mathcal{K}^i = \begin{bmatrix} K_1(1,i) & \cdots & K_M(1,i) \\ \vdots & \ddots & \vdots \\ K_1(N,i) & \cdots & K_M(N,i) \end{bmatrix} \in \mathbb{R}^{N \times M}$$

Bài toán trên dễ ràng mở rộng ra trong trường hợp không gian mới với số chiều > 1 chiều. Khi đó ma trận chiếu $A = [\alpha_1, \alpha_2, ...\alpha_P]$ trong đó P là số chiều trong không gian mới và mỗi vector α_i là một vector chiếu trong chiều tương ứng.

Giải bài toán (3) dựa trên việc đồng thời phải tối ưu cả 2 hệ số A và β. Việc giải bài toán tối ưu trên cả 2 nghiệm là rất khó, chính vì vậy, một giải pháp được đưa ra là tối ưu hóa lần lượt trên từng nghiệm. Nghĩa là, tại mỗi lần lặp ta sẽ cố định A để giải bài toán tối ưu trên một nghiệm duy nhất là β, sau đó lại cố định β để giải bài toán tối ưu trên một nghiệm duy nhất là A. Việc khởi tạo β và A cho lần lặp đầu tiên có thể thực hiện như sau: Gán tất cả các giá trị của các phần tử trong vector trọng số β tương ứng với trọng số của các nhân đều bằng 1, hoặc khởi tạo ma trận $AA^T=I$.

Chúng tôi sử dụng nền tảng MKL-DR kết hợp với thuật toán giảm chiều dữ liệu là "Phép chiếu bảo toàn tính cục bộ - Locality Preserving Projections (LPP)" (He và Niyogi, 2004). Đây là một phương pháp học không giám sát nhằm bảo toàn khoảng cách của một điểm dữ liệu đến k điểm dữ liệu gần nó nhất. Với LPP, ma trận W và D được xác định như sau:

$$w_{ij} = \begin{cases} 1, \text{nếu } i \in \mathcal{N}_k(j) \vee j \in \mathcal{N}_k(i) \\ 0, \text{các trường hợp còn lại} \end{cases} \qquad d_{ij} = \begin{cases} \sum_{n=1}^{N} w_{in}, \text{nếu } i = j \\ 0, \text{các trường hợp còn lại} \end{cases}$$

Việc phân cụm được thực hiện bởi thuật toán k-means. Để đánh giá kết quả phân cụm chúng tôi sử dụng phương pháp của Rousseeuw (1987), đây là một độ đo để đánh giá xem một điểm dữ liệu phù hợp ở mức độ nào với cụm của nó và nó phù hợp như thế nào trong các cụm khác. Khi lấy trung bình trên tất cả các điểm dữ liệu, kết quả giá trị hình chiếu trung bình sẽ cho biết các cụm được phân tách tốt như thế nào.

Kết quả và thảo luận

*** Dữ liệu sử dụng**

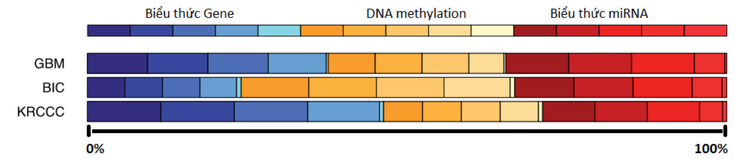
Chúng tôi áp dụng MKL-DR cho 3 tập dữ liệu ung thư được lấy từ TCGA (The Cancer Genome Atlats) là: Glioblastoma Multiforme (GBM) với 213 bệnh nhân, Breast Invasive Carcinoma (BIC) với 105 bệnh nhân và Kidney Renal Clear Cell Carcinoma (KRCCC) với 122 bệnh nhân, mới mỗi loại ung thư, chúng tôi sử dụng biểu thức gen, DNA methylation và biểu thức miRNA để tiến hành phân cụm. Với mỗi loại biểu diễn chúng tôi xây dựng 5 ma trận nhân tương ứng sử dụng hàm nhân cơ sở Gaussian với hệ số $= 1/(2\delta^2)$ với δ là số chiều của dữ liệu và các hệ số tương ứng với 5 ma trận nhân $\gamma = \{10^{-6}\gamma_1, 10^{-3}\gamma_1, \gamma_1, 10^3\gamma_1, 10^6\gamma_1\}$.

Với mỗi tập dữ liệu chúng tôi sử dụng cả 2 cách khởi tạo lần lượt với cả A và β, sau đó từ dữ liệu đã được tích hợp chúng tôi tiến hành phân cụm với thuật toán k-means với k ∈ {2,...,15} sau đó chọn ra số cụm tối ưu sử dụng giá trị hình chiếu trung bình của kết quả phân cụm. Với tham số đầu vào khi sử dụng thuật toán LPP thì số điểm dữ liệu gần nhất được gán bằng 9.

*** Kết quả thu được như sau:**

- Đóng góp của các loại biểu diễn đặc trưng cho dữ liệu được tích hợp:

Hình 1 thể hiện sự đóng góp của từng loại biểu diễn đặc trưng cho dữ liệu được tích hợp cuối cùng. Đóng góp của từng loại dữ liệu cho dữ liệu tích hợp là khác nhau, với GBM và KRCCC đóng góp của DNA methylation ít hơn so với biểu thức Gene và miRNA, còn đối với BIC thì đóng góp của biểu thức Gene ít hơn đáng kể so với DNA methylation và biểu thức miRNA.



Hình 1. Đóng góp của các loại biểu diễn dữ liệu cho dữ liệu được tích hợp

- Kết quả phân cụm

So sánh kết quả phân cụm với phương pháp Similarity Network Fusion - SNF của (Wang và các đồng sự, 2014) bằng cách xem xét giá trị P cho mô hình hồi quy Cox (Hosmer và các đồng sự, 2011) và thu được kết quả thể hiện qua bảng sau:

Loại ung thư	SNF	MKL-LPP
GBM	2.0E-4 (3)	6.5E-6 (6)
BIC	1.1E-3 (5)	3.4E-3 (7)
KRCCC	2.9E-2 (3)	4.0E-5 (14)

Bảng 1. So sánh kết quả phân cụm với phương pháp SNF

Lưu ý: Số trong ngoặc là số cụm

Kết quả cho thấy giá trị trung vị của SNF là 1.1E-3 và của MKL-LPP là 4.0E-5 cho thấy hiệu năng của MKL-LPP tốt hơn so với SNF và là một mô hình đáng tin cậy để phân cụm đối tượng.

Kết luận

Từ kết quả trên cho thấy đây là một phương pháp hiệu quả để thực hiện phân cụm dữ liệu bệnh nhân ung thư, từ kết quả phân cụm có thể tiến hành phân tích để đề xuất các giải pháp cho chuẩn đoán và điều trị bệnh ung thư.

Trong thời gian tới, chúng tôi đang nghiên cứu để áp dụng phương pháp MKL-DR có thể xây dựng một mô hình phân lớp dự đoán bệnh ung thư.

Tài liệu tham khảo

[1] He,X. and Niyogi,P. (2004) Locality preserving projections. In: Thrun,S. et al (eds.) *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, pp. 153–160.

[2] Yan,S. et al. (2007) Graph embedding and extensions: a general framework for dimensionality reduction. *IEEE Trans. Pattern Anal. Machine Intell.*, 29, 40–51.

[3] Lin,Y.-Y. et al. (2011) Multiple kernel learning for dimensionality reduction. *IEEE Trans. Pattern Anal. Machine Intell.*, 33, 1147–1160.

[4] Hosmer,D.W., Jr. et al. (2011) Applied Survival Analysis: Regression Modeling of Time to Event Data. Wiley, New York.

[5] Wang,B. et al. (2014) Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods*, 11, 333–337.

[6] Nora K. Speicher and Nico Pfeifer. (2015) Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics*, 31, 2015, i268–i275