SO SÁNH MỘT SỐ THUẬT TOÁN PHÂN CỤM PHỔ CHO DỮ LIỆU BIỂU DIỄN GENE

Hoàng Thị Thanh Giang*, Nguyễn Thị Thúy Hạnh, Nguyễn Hoàng Huy

Khoa Công nghệ Thông tin, Học viện Nông nghiệp Việt Nam

Email: httgiang@vnua.edu.vn*

Ngày gửi bài: 22.07.2015 Ngày chấp nhận: 03.09.2015

TÓM TẮT

Các thuật toán phân cụm phổ là một trong những thuật toán hiệu quả nhất để phân chia các gene thành các nhóm theo mức độ tương tự biểu diễn gene của chúng. Những phân nhóm như thế có thể đề xuất những gene tương ứng tương quan và/hoặc cùng được điều hòa và dẫn đến chỉ ra những gene đó có thể chia sẻ một vai trò sinh học chung. Trong bài báo này, ba thuật toán phân cụm phổ phổ biến nhất được nghiên cứu: phân cụm phổ không chuẩn hóa, phân cụm phổ chuẩn hóa theo Shi và Malik (2000), phân cụm phổ chuẩn hóa theo Ng et al. (2002). Những thuật toán này được so sánh với nhau. Hiệu năng của ba thuật toán này được nghiên cứu trên dữ liệu chuỗi thời gian của biểu diễn gene sử dụng khoảng cách xoắn thời gian động (DTW) để đo độ tương tự giữa những hồ sơ thể hiện gene. Bốn độ đo hiệu lực phân cụm khác nhau được sử dụng để đánh giá các thuật toán phân cụm: Độ đo liên kết (Connectivity) và chỉ số Silhouette (Silhouette Index) để ước lượng chất lượng của phân cụm, chỉ số Jaccard (Jaccard Index) để đánh giá độ ổn định của phương pháp phân cụm và chỉ số Rand (Rand Index) để đánh giá sự chính xác. Sau đó chúng tôi phân tích các kết quả thu được bởi kiểm định Friedman. Phân cụm phổ chuẩn hóa theo Ng et al. (2002) chứng tỏ là tốt nhất theo chỉ số hiệu lực Silhouette và Rand.

Từ khóa: Hồ sơ biểu diễn gene, phân cụm phổ chuẩn hóa, phân cụm phổ không chuẩn hóa.

Comparison of Spectral Clustering Algorithms for Gene Expression Data

ABSTRACT

Spectral clustering algorithms have been the most effective algorithms to divide genes into groups according to the degree of their expression similarity. Such a grouping may suggest that the respective genes are correlated and/or co-regulated, and subsequently indicates that the genes could possibly share a common biological role. In this paper, three spectral clustering algorithms were investigated: unnormalized spectral clustering, normalized spectral clustering according to Shi and Malik (2000), and normalized spectral clustering according to Ng, Jordan and Weiss (2002). The algorithms were benchmarked against each other. The performance of the three clustering algorithms was studied on time series expression data using Dynamic Time Warping (DTW) distance in order to measure similarity between gene expression profiles. Four different cluster validation measures were used to evaluate the clustering algorithms: Connectivity and Silhouette Index for estimatingthe quality of clusters, Jaccard Index for evaluating the stability of a cluster method and Rand Index for assessing the accuracy. The results were analyzed by Friedman's test. The performance of normalized spectral clustering according to Ng, Jordan and Weiss (2002) was demonstrated to be the best under the Silhouette and Rand validation indices.

Keywords: Normalized spectral clustering, unnormalized spectral clustering, gene expression profiles.

1. ĐẶT VẤN ĐỀ

Vi mạng biểu diễn gene là nguồn sẵn có nhất của dữ liệu sinh học thông lượng cao. Mỗi thí nghiệm vi mạng đo mức độ biểu diễn của một tập hợp các gene trong những điều kiện thí nghiệm hoặc tại những điểm thời gian khác nhau. Sự tiến bộ trong kỹ thuật phân tích gene dẫn đến sự gia tăng tập dữ liệu biểu diễn gene cả về kích cỡ và số lượng (Datta, 2003; Xu et al., 2002). Phân cụm gene là một trong những nhiệm vụ phân tích vi mạng quan trọng nhất

trong trích xuất những thông tin có nghĩa từ hồ sơ biểu diễn gene. Ví dụ, một sự tương tự cao giữa những hồ sơ biểu diễn gene có thể gợi ý rằng những gene tương ứng tương quan và/hoặc cùng được điều hòa, rồi dẫn đến chỉ ra những gene này có thể chia sẻ một vai trò sinh học chung. Vì thế nhóm các gene theo sự tương tự biểu diễn của chúng có thể làm gia tăng sự hiểu biết của những chức năng gene, quá trình tế bào và sự liên hệ giữa các gene (Datta, 2003; Quackenbush, 2001; Jiang, 2004)

Những phương pháp phân cụm thông dụng nhất như phân cụm cấp bậc thường trực quan và dễ sử dụng, tuy nhiên chúng yêu cầu những sự lựa chọn tùy hứng trên các tham số khác nhau (như ngưỡng để cắt cây ...). Dữ liệu biểu diễn gene nói chung khó phân cụm một cách hiệu quả do sự đa dạng của kiểu gene (Huang et al., 2013). Hơn nữa, sự lựa chọn của các thuật toán phân cụm phụ thuộc vào dữ liệu được khám phá. Chất lượng của lời giải phân cụm cũng bị ảnh hưởng bởi độ đo sử dụng để đánh giá sự tương tự (khoảng cách) giữa các hồ sơ biểu diễn gene (Borg et al., 2013). Tác giả Huang et al. (2013) đã chỉ ra rằng một số thuật toán phân cụm phổ tốt hơn những phương pháp cấp bậc và tốt bằng những thuật toán k – means, nhưng nhanh hơn chúng, cho phân cụm dữ liệu biểu diễn gene không đồng nhất.

Trong bài báo này, chúng tôi nghiên cứu ba thuật toán phân cụm phổ phổ biến nhất để phân chia dữ liệu vi mảng DNA. Những thuật toán này được đánh giá và so sánh với nhau trên những chuỗi thời gian của biểu diễn gene thu được từ một nghiên cứu kiểm tra sự điều hòa chu kỳ tế bào toàn thể trong sự phân hạch nấm men *Schizosaccharomyces pombe*. Chuỗi thời gian của những hồ sơ biểu diễn gene được cho rằng không chỉ biến đổi về biên độ biểu diễn mà còn về sự tiến triển theo thời gian do quá trình sinh học có thể bộc lộ ra với tốc độ khác nhau tương ứng với điều kiện thí nghiệm khác nhau hoặc bên trong những cơ quan và cá thể khác nhau. Do vậy, những khoảng cách metric cổ điển như Ơclit, Manhattan,... sẽ đưa ra điểm số tương tự nghèo nàn. Bởi lý do này, những thuật toán được nghiên cứu sử dụng khoảng cách xoắn thời gian động (Dynamic Time Warping) để đo độ tương tự giữa những hồ sơ biểu diễn gene. Bốn độ đo hiệu lực của phân cụm được sử dụng để đánh giá hiệu năng của các thuật toán, cho phép đánh giá và so sánh những khía cạnh khác nhau của lời giải phân cụm có được. Hơn nữa kiểm định thống kê được áp dụng để xác định sự khác biệt giữa các thuật toán.

Phần còn lại được sắp xếp như sau. Phần 2 giới thiệu về các thuật toán phân cụm phổ. Phần 3 trình bày về thiết lập thí nghiệm. Phần 4 đưa ra những kết quả và thảo luận. Phần 5 kết luận bài báo.

2. VẬT LIỆU VÀ PHƯƠNG PHÁP

2.1. Phương pháp nghiên cứu

2.1.1. Đồ thị tương tự

Cho dữ liệu gồm n chuỗi thời gian $x_1, \cdots, x_n$ biểu diễn của n gene. Ở đây, chúng ta nghiên cứu phương pháp phân cụm để xác định các mẫu của biểu diễn gene, với mục đích tăng sự hiểu biết về chức năng của biểu diễn gene hoặc mối liên hệ giữa biểu diễn gene (Quackenbush, 2001; Chen et al., 2007). Những mẫu này có thể được khám phá ra dựa trên độ tương tự, hoặc khoảng cách giữa các cặp hồ sơ biểu diễn gene (Nguyen and Li, 2009). Hai trong số các hàm khoảng cách được sử dụng phổ biến trong thực tế là khoảng cách Ơclit, tương quan Pearson (Quackenbush, 2001). Tuy nhiên chúng thích hợp khi so sánh những chuỗi thời gian biểu diễn gene vì chuỗi thời gian biểu diễn gene có thể tiến triển ở những tốc độ khác nhau. Để làm ngang bằng sự khác nhau trong tốc độ tiến triển, khoảng cách xoắn thời gian động (Dynamic Time Warping Distance) có thể được sử dụng (Al – Naymat et al., 2009). Sau đây chúng ta ký hiệu khoảng cách xoắn thời gian động của cặp gene (x_i, x_j) là $d_{ij} = DTW(x_i, x_j)$.

Xây dựng đồ thị tương tự cho phân cụm phổ không phải là nhiệm vụ dễ dàng, do có rất ít cơ sở lý thuyết. Có một số cách xây dựng phổ biến để biến đổi n chuỗi thời gian $x_1, \cdots, x_n$ của n gene đã cho với khoảng cách của từng cặp d_{ij} thành một đồ thị như đồ thị ϵ – lân cận, đồ thị k – hàng

xóm gần nhất, đồ thị liên thông đầy đủ (Luxburg, 2007). Ở đây chúng tôi nghiên cứu các thuật toán phân cụm phổ sử dụng đồ thị k - hàng xóm gần nhất. Tuy nhiên để có thể sử dụng hai phương pháp xây dựng đồ thị tương tự trên, trước tiên chúng ta phải xây dựng đồ thị tương tự đầy đủ sử dụng khoảng cách xoắn thời gian động. Trong đồ thị này mỗi đỉnh x_i đại diện cho mỗi gene $x_i, i = 1, \cdots, n$, trong đó n là tổng số gene và các cạnh có trọng số không âm đối xứng s_{ij} mã hóa độ ảnh hưởng lẫn nhau giữa các cặp gene. Độ ảnh hưởng biểu thị khả năng một cặp gene thuộc vào cùng nhóm. Ở đây chúng tôi sử dụng một dạng biến đổi của khoảng cách xoắn thời gian động d_{ij}như là độ ảnh hưởng, được tính trên chuỗi thời gian biểu diễn gene.

$$s_{ij} = exp\left(-\left(sin\frac{arccos(d_{ij}/d_{max})}{2}\right)^2\right)$$

trong đó $d_{max} = max\{|d_{ij}|, i, j = 1, \cdots, n\}$. Độ tương tự này đã đem lại thành công thực nghiệm trong phân cụm dữ liệu biểu diễn gene (Frey and Dueck, 2007).

Trong cách xây dựng đồ thị ϵ - lân cận, chúng tôi nối tất cả các gene mà khoảng cách giữa chúng nhỏ hơn ϵ. Khi khoảng cách giữa các đỉnh kề nhau xấp xỉ cùng một mức (tối đa ϵ), đánh trọng số các cạnh không thể kết hợp thêm thông tin từ dữ liệu vào đồ thị. Do đó đồ thị ϵ - lân cận thường được coi là một đồ thị không trọng số. Đối với đồ thị ϵ - lân cận, chúng ta có thể thấy rất khó để chọn một tham số ϵ hữu ích. Điều này thường xảy ra nếu chúng ta có dữ liệu trên những thang khác nhau, do đó khoảng cách giữa các điểm dữ liệu là khác nhau trong những vùng khác nhau của không gian.

Mục đích của đồ thị k - hàng xóm gần nhất là nối đỉnh (gene) x_i với đỉnh x_j nếu x_j nằm trong k - hàng xóm gần nhất của x_i. Tuy nhiên, định nghĩa này dẫn đến một đồ thị định hướng vì quan hệ hàng xóm không đối xứng. Có hai cách để làm đồ thị này vô hướng. Cách đầu tiên đơn giản là bỏ qua hướng của cạnh, đó là chúng ta nối x_i và x_j với một cạnh không định hướng nếu x_i nằm trong k - hàng xóm gần nhất của x_j hoặc nếu x_j nằm trong k - hàng xóm gần nhất của x_i. Đồ thị thu được thường được gọi là đồ thị k - hàng xóm gần nhất. Lựa chọn thứ hai là nối hai đỉnh x_i và x_j nếu cả x_i nằm trong k hàng xóm gần nhất của x_j và x_j nằm trong k hàng xóm gần nhất của x_i. Đồ thị thu được gọi là đồ thị k - hàng xóm gần nhất chung. Trong cả hai trường hợp, sau khi nối những đỉnh thích hợp chúng ta đánh trọng số của cạnh bằng độ tương tự giữa các đỉnh. Khác với đồ thị ϵ - lân cận, đồ thị k - hàng xóm gần nhất có thể kết nối các điểm dữ liệu trên các thang khác nhau. Đây là một tính chất tổng quát rất hữu ích của đồ thị k - hàng xóm gần nhất. Đồ thị k - hàng xóm gần nhất có thể cắt thành một vài thành phần không liên thông nếu có một vài vùng mật độ cao tương đối xa nhau. Đồ thị k - hàng xóm gần nhất chung có xu hướng nối điểm dữ liệu trong những vùng có mật độ như nhau, nhưng không nối những vùng có mật độ khác nhau. Do đó đồ thị k - hàng xóm gần nhất chung có thể coi như "nằm giữa" đồ thị ϵ - lân cận và đồ thị k - hàng xóm gần nhất. Nó có thể tác động trên những thang khác nhau, nhưng không trộn lẫn những thang này với nhau. Ký hiệu $S = (s_{ij})$ là ma trận trọng số của đồ thị. Hơn nữa, với đồ thị tương tự đầy đủ, ma trận trọng số S không là ma trận thưa nhưng với đồ thị k - hàng xóm gần nhất chung, ma trận trọng số S là ma trận thưa. Do đó chúng tôi sử dụng đồ thị k - hàng xóm gần nhất chung khi ứng dụng các thuật toán phân cụm phổ cho dữ liệu biểu diễn gene.

2.1.2. Đồ thị Laplace

Công cụ chính cho phân cụm phổ là các ma trận đồ thị Laplace. Tồn tại cả một lĩnh vực dành cho nghiên cứu các ma trận này, gọi là lý thuyết đồ thị phổ (Chung, 1997). Trong phần này chúng tôi muốn định nghĩa những ma trận đồ thị Laplace khác nhau. Chú ý rằng trong tài liệu, không có quy tắc duy nhất gọi tên các ma trận đồ thị Laplace. Thông thường, mỗi tác giả chỉ gọi ma trận của "anh ta" là ma trận Laplace. Do đó cần thận trọng khi đọc tài liệu về ma trận đồ thị Laplace.

Sau đây chúng ta giả sử rằng $G = (V, E)$ là một đồ thị không định hướng trên các đỉnh $= \{_1, \cdots, \}$. Trong những phần tiếp theo chúng ta giả sử rằng đồ thị G có trọng số, mỗi cạnh giữa đỉnh và mang một trọng số không âm ≥ 0. Ma trận trọng số của đồ thị là ma trận $S = (s_{ij})_{i,j=1,\ldots,n}$, $s_{ij} = 0$, nghĩa là đỉnh x_i và x_j không nối nhau. Vì G là không định hướng, chúng ta có $s_{ij} = s_{ji} \geq 0$. Khi sử dụng vectơ riêng của một ma trận, chúng ta sẽ không cần giả sử rằng chúng được chuẩn hóa. Ví dụ, vectơ hằng $\mathbf{I}$ và bội $a\mathbf{I}$ với $a \neq 0$ sẽ được coi như các vectơ riêng giống nhau. Các giá trị riêng sẽ luôn được sắp xếp tăng dần tương ứng với số bội. "k vectơ riêng đầu tiên" được quy cho những vectơ riêng tương ứng với k giá trị riêng nhỏ nhất.

Ma trận của đồ thị Laplace không chuẩn hóa được xác định bởi Mohar (1991, 1997):

$$L = D - S.$$

trong đó, D là ma trận chéo với các phần tử chéo là các bậc của các đỉnh $x_i, i = 1, \cdots, n$

$$d_i = \sum_{j=1}^{n} s_{ij}.$$

Chú ý rằng, thật sự tổng này chỉ chạy trên tất cả các cạnh liền kề với x_i, vì đối với tất cả những đỉnh x_j còn lại trọng số $s_{ij} = 0$.

Có hai cách định nghĩa ma trận của đồ thị Laplace chuẩn hóa trong các tài liệu tham khảo. Cả hai ma trận này quan hệ mật thiết với nhau và được định nghĩa như sau (Chung, 1997):

$$L_{sym} \coloneqq D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} S D^{-\frac{1}{2}}$$

$$L_{rw} \coloneqq D^{-1} L = I - D^{-1} S.$$

Chúng ta ký hiệu ma trận đầu tiên là L_{sym} bởi nó là một ma trận đối xứng, ma trận thứ hai là L_{rw} bởi nó liên hệ mật thiết với phương pháp bước đi ngẫu nhiên (random walk).

Bây giờ chúng tôi xin phát biểu ba thuật toán phân cụm phổ phổ biến nhất: phân cụm phổ không chuẩn hóa, phân cụm phổ chuẩn hóa theo Shi and Malik (2000), phân cụm phổ chuẩn hóa theo Ng et al. (2002). Chúng ta giả sử có dữ liệu gồm n chuỗi thời gian $\{x_1, \ldots, x_n\}$ của n gene. Chúng ta đo độ tương tự của từng cặp gene $s_{ij} = s(x_i, x_j)$ bằng phương pháp đã miêu tả trong phần 2.1.1. và ký hiệu ma trận tương tự tương ứng bởi $S = (s_{ij})_{i,j=1,\ldots,n}$.

2.1.3. Phân cụm phổ không chuẩn hóa

Đầu vào: Ma trận tương tự $S = (s_{ij})_{i,j=1\cdots n} \in \mathbb{R}^{n \times n}$, số cụm k để xây dựng

- Xây dựng một đồ thị tương tự bằng một trong những cách được miêu tả trong phần 2.1.1.. Đặt S là ma trận trọng số của nó.
- Tính ma trận Laplace không chuẩn hóa L.
- **Tính k vectơ riêng đầu tiên $u_1, \cdots, u_k$ của ma trận L.**
- Đặt $U \in \mathbb{R}^{n \times k}$ là ma trận gồm các cột là các vectơ $u_1, \cdots, u_k$.
- Với $i = 1, \cdots, n$, đặt $y_i \in \mathbb{R}^k$ là vectơ tương ứng với hàng thứ i của U.
- Phân cụm các điểm $(y_i)_{i=1,\ldots,n}$ trong $\mathbb{R}^k$ với thuật toán k - means thành các cụm $\mathsf{C}_1, \ldots, \mathsf{C}_k$.

Đầu ra: Các cụm $A_1, \ldots, A_k$ với $A_i = \{j | y_j \in \mathsf{C}_i\}$.

Có hai phiên bản khác nhau của phân cụm phổ chuẩn hóa, phụ thuộc đồ thị Laplace chuẩn hóa được sử dụng. Tên của hai thuật toán được đặt theo hai bài báo phổ biến (Shi and Malik, 2000; Ng et al., 2002).

2.1.4. Phân cụm phổ chuẩn hóa theo Shi and Malik (2000)

Đầu vào: Ma trận tương tự $S = (s_{ij})_{i,j=1\cdots n} \in \mathbb{R}^{n \times n}$, số cụm k để xây dựng

- Xây dựng một đồ thị tương tự bằng một trong những cách được miêu tả trong phần 2.1.1.. Đặt S là ma trận trọng số của nó.
- Tính ma trận Laplace không chuẩn hóa L.
- **Tính k vectơ riêng suy rộng đầu tiên $u_1, \cdots, u_k$ của bài toán riêng suy rộng $Lu = \lambda Du$.**
- Đặt $U \in \mathbb{R}^{n \times k}$ là ma trận gồm các cột là các vectơ $u_1, \cdots, u_k$.
- Với $i = 1, \cdots, n$, đặt $y_i \in \mathbb{R}^k$ là vectơ tương ứng với hàng thứ i của U.

- Phân cụm các điểm $(y_i)_{i=1,\dots,n}$ trong $\mathbb{R}^k$ với thuật toán k - means thành các cụm $C_1, \dots, C_k$.

Đầu ra: Các cụm $A_1, \dots, A_k$ với $A_i = \{j | y_j \in C_i\}$.

Chú ý rằng thuật toán này sử dụng các vectơ riêng suy rộng của L, theo (Chung, 1997) chúng tương đương với các vectơ riêng của ma trận L_{rw}. Vì vậy, thực tế thuật toán làm việc với các vectơ riêng của ma trận Laplace chuẩn hóa L_{rw} và do đó được gọi là phân cụm phổ chuẩn hóa. Thuật toán tiếp theo cũng sử dụng một ma trận Laplace chuẩn hóa, nhưng lần này là ma trận L_{sym} thay cho L_{rw}. Như chúng ta sẽ thấy, thuật toán này cần đưa thêm bước chuẩn hóa hàng, bước này không cần trong những thuật toán phổ khác. Lý do vì sao được trình bày rõ ràng trong bài báo (Ng et al., 2002).

2.1.5. Phân cụm phổ chuẩn hóa theo Ng et al. (2002)

Đầu vào: Ma trận tương tự $S = (s_{ij})_{i,j=1\cdots n} \in \mathbb{R}^{n\times n}$, số cụm k để xây dựng

- Xây dựng một đồ thị tương tự bằng một trong những cách được miêu tả trong phần 2.1.1. Đặt S là ma trận trọng số của nó.
- Tính ma trận Laplace không chuẩn hóa L.
- **Tính k vectơ riêng suy rộng đầu tiên $u_1, \cdots, u_k$ của ma trận L_{sym}.**
- Đặt $U \in \mathbb{R}^{n\times k}$ là ma trận gồm các cột là các vectơ $u_1, \cdots, u_k$.
- Với $i = 1, \cdots, n$, đặt $y_i \in \mathbb{R}^k$ là vectơ tương ứng với hàng thứ i của U.
- **Thành lập ma trận $T \in \mathbb{R}^{n\times k}$ từ U bằng cách chuẩn hóa các hàng sao cho các hàng có chuẩn 1, nghĩa là các phần tử của ma trận T xác định bởi $t_{ij} = \frac{u_{ij}}{(\sum_k u_{ik}^2)^{\frac{1}{2}}}$.**
- Phân cụm các điểm $(y_i)_{i=1,\dots,n}$ trong $\mathbb{R}^k$ với thuật toán k - means thành các cụm $C_1, \dots, C_k$.

Đầu ra: Các cụm $A_1, \dots, A_k$ với $A_i = \{j | y_j \in C_i\}$.

Tất cả ba thuật toán phát biểu ở trên trông khá giống nhau, ngoại trừ việc chúng sử dụng ba đồ thị Laplace khác nhau. Trong cả ba thuật toán, thủ thuật chính là thay sự biểu diễn của các điểm dữ liệu trừu tượng x_i thành điểm $y_i \in \mathbb{R}^k$. Do các tính chất của đồ thị Laplace, những thay đổi biểu diễn làm tăng khả năng phân cụm trong dữ liệu, đến mức các cụm có thể được phát hiện dễ dàng trong biểu diễn mới. Cụ thể, thuật toán phân cụm k - means không có khó khăn để phát hiện các cụm trong biểu diễn mới.

2.2. Vật liệu nghiên cứu

Trong mục này, đầu tiên chúng tôi mô tả tập dữ liệu vi mảng được sử dụng để chứng thực và đánh giá các thuật toán phân cụm được miêu tả trong mục trên. Tiếp theo, chúng tôi cung cấp một tổng quan ngắn gọn của những độ đo hiệu lực của phân cụm được sử dụng.

2.2.1. Tập dữ liệu vi mảng

Kết quả phân cụm của những thuật toán phân cụm kể trên được đánh giá trên dữ liệu chuỗi thời gian biểu diễn gene thu được từ một nghiên cứu sự điều hòa chu kỳ tế bào toàn thể trong sự phân hạch nấm men *Schizosaccharomyces pombe* (Rustici et al., 2004). Nghiên cứu này bao gồm 8 thí nghiệm độc lập theo tiến trình thời gian đồng bộ hóa bởi: 1) Sự gạn sạch (lặp sinh học độc lập 3 lần); 2) Giải phóng khối cdc25 (lặp sinh học độc lập 2 lần, một trong hai sao chép kỹ thuật đổi nhuộm và một thí nghiệm trong nền đột biến sep1) và 3) Một phối hợp của hai phương pháp (gạn sạch và giải phóng khối cdc25 cũng như gạn sạch và giải phóng khối cdc10). Do đó, có chín tập dữ liệu biểu diễn gene kiểm tra khác nhau: elu1, elu2, elu3, cdc25 - 1, cdc25 - 2.1, cdc25 - 2.2, cdc25 - sep1, elu - cdc10 và elu - cdc25. Trong pha tiền xử lý số liệu những dòng với thiếu hơn 25% số phần tử bị lọc khỏi mỗi ma trận biểu diễn và những phần tử biểu diễn còn thiếu khác được đưa vào bằng thuật toán DTWinpute (Tsiporkova and Boeva, 2007). Bằng cách này, chúng ta thu được 9 ma trận đầy đủ.

Rustici et al. (2004) đã xác định 407 gene có chức năng điều hòa chu kỳ tế bào. Chúng được đưa ra để phân cụm điều này dẫn đến hình thành bốn cụm rời nhau. Những gene không xuất hiện cùng trong 9 tập dữ liệu gốc đã được

loại bỏ, còn lại 267 gene. Rồi sau đó, hồ sơ biểu diễn theo thời gian của những gene này được trích rút từ ma trận dữ liệu đầy đủ, hình thành tập dữ liệu tiêu chuẩn gồm 9 ma trận mới.

Các tập dữ liệu tiêu chuẩn này được chuẩn hóa thêm nữa bằng cách áp dụng một phương pháp biến đổi dữ liệu được đưa ra bởi Boeva and Tsiporkova (2010).

2.2.2. Độ đo hiệu lực phân cụm

Một trong những vấn đề quan trọng nhất của phân tích phân cụm là hiệu lực phân cụm. Về bản chất những kỹ thuật đánh giá hiệu lực phân cụm được thiết kế để tìm sự phân chia khớp nhất với các cụm đã biết và do đó nên được coi là một chìa khóa để hiểu kết quả phân cụm.

Do không một thuật toán phân cụm nào đều thể hiện tốt nhất trong tất cả các kịch bản, nên sử dụng một độ đo hiệu lực đơn lẻ là không đáng tin cậy, do vậy thay vào đó chúng tôi sử dụng ít nhất hai độ đo phản ánh những khía cạnh khác nhau của một sự phân chia. Với cách nghĩ này, chúng tôi thực hiện bốn độ đo hiệu lực khác nhau. Để đánh giá chất lượng của phân cụm chúng tôi sử dụng **độ đo liên kết** (Connectivity) để ước lượng sự kết nối (Handl et al., 2005) và **chỉ số Silhouette** (Silhouette Index) cho ước lượng tính chia cắt và chặt của một sự phân chia. Hơn nữa **chỉ số Jaccard** (Jaccard Index) được sử dụng cho đánh giá sự ổn định của một phương pháp phân cụm (Jaccard, 1912). Phương pháp xem xét được ngẫu nhiên hóa, sao cho khi áp dụng p lần cho các kết quả khác nhau. Chúng tôi tính chỉ số Jaccard trung bình qua tất cả $p(p - 1)/2$ cặp của p kết quả, sự ổn định của phương pháp được ước lượng bởi chỉ số này. Cuối cùng chúng tôi sử dụng **chỉ số Rand** (Rand Index) cho ước lượng độ chính xác (Rand, 1971). Độ đo này được áp dụng để tính toán sự tương hợp giữa kết quả phân cụm sinh bởi phương pháp được xem xét và lời giải phân cụm đã biết (lời giải đúng) (Rustici et al., 2004).

3. KẾT QUẢ VÀ THẢO LUẬN

Trong phần này, hiệu năng của ba thuật toán phân cụm phổ được nghiên cứu trên ma trận biểu diễn gene chuẩn bằng cách sử dụng độ đo hiệu lực phân cụm được miêu tả trong phần 2.2.2. Kiểm định Friedman được áp dụng để phát hiện sự khác biệt giữa các thuật toán.

3.1. Chất lượng của phân cụm

Trong phần này, chúng tôi đánh giá và so sánh chất lượng của các lời giải phân cụm được sinh ra bởi ba thuật toán phân cụm phổ được thảo luận trong phần 2.1. trên tập dữ liệu chuẩn được miêu tả trong phần 2.2.1. bằng cách sử dụng hai độ đo hiệu lực phân cụm: chỉ số Silhouette và độ đo liên kết. Trong bảng 1(a), các hàng biểu diễn giá trị trung bình độ đo liên kết của các lời giải phân cụm thu được bởi các thuật toán phân cụm phổ khác nhau trên tập dữ liệu chuẩn. Bậc trung bình của các thuật toán chỉ ra thuật toán phân cụm phổ chuẩn hóa theo Shi and Malik (2000) có hiệu năng tốt nhất. Kiểm định Friedman chỉ ra rằng có sự khác biệt giữa các thuật toán, $X^2 = 9{,}1, df = 2, p = 0{,}05$. Tương tự, bảng 1(b) biểu diễn điểm trung bình của chỉ số Silhouette. Bậc trung bình của các thuật toán chỉ ra thuật toán phân cụm phổ chuẩn hóa theo Ng et al. (2002) có hiệu năng tốt nhất. Kiểm định Friedman chỉ ra rằng có khác biệt giữa các thuật toán, $X^2 = 10{,}9, df = 2, p = 0{,}01$.

3.2. Sự ổn định phân cụm

Trong phần này, chúng tôi đánh giá và so sánh sự ổn định của các lời giải phân cụm tạo ra bằng cách sử dụng chỉ số Jaccard. Như có thể thấy từ bảng 1(c), thuật toán phân cụm phổ không chuẩn hóa có hiệu năng tốt nhất theo chỉ số Jaccard. Nhìn chung các thuật toán chứng tỏ có sự ổn định tốt. Theo kiểm định Friedman, có sự khác biệt tồn tại giữa các thuật toán, $X^2 = 10{,}4, df = 2, p = 0{,}01$.

3.3. Độ phân cụm chính xác

Trong phần này, chỉ số Rand được sử dụng để đánh giá độ chính xác của lời giải phân cụm tạo bởi các thuật toán kể trên. Như có thể thấy trong bảng 1(d), thuật toán phân cụm phổ không chuẩn hóa và chuẩn hóa theo Shi and Malik (2000) chứng tỏ độ chính xác tương đối thấp

Bảng 1. Độ đo phân cụm trung bình và bậc thuật toán trung bình

a. Độ đo liên kết

	d1	d2	d3	d4	d5	d6	d7	d8	d9	R^a
kch	199,2	162,8	183,7	177,8	172,2	182,7	155,9	104	162,9	2,92
sm	66,8	25	45,5	6	4	4	5	14	12	1,22
njw	123	159,6	186,2	166	166	166	159,6	166	139,8	2,22

b. Chỉ số Silhouette

	d1	d2	d3	d4	d5	d6	d7	d8	d9	R^a
kch	0,54	0,51	0,38	0,42	0,39	0,40	0,38	0,56	0,55	2,88
sm	- 0,49	- 0,29	- 0,21	0,83	0,84	0,95	- 0,48	0,58	0,62	2,88
njw	0,55	0,54	0,44	0,47	0,32	0,42	0,36	0,65	0,57	1,44

c. Chỉ số Jaccard

	d1	d2	d3	d4	d5	d6	d7	d8	d9	R^a
kch	1,00	1,00	0,85	1,00	0,91	0,71	0,86	1,00	1,00	1,28
sm	0,74	0,70	0,72	0,92	0,94	0,94	0,93	0,82	0,84	2,33
njw	0,99	0,99	0,79	0,65	0,65	0,65	0,99	0,53	0,43	2,88

d. Chỉ số Rand

	d1	d2	d3	d4	d5	d6	d7	d8	d9	R^a
kch	0,43	0,44	0,43	0,36	0,35	0,35	0,36	0,39	0,38	2,91
sm	0,34	0,34	0,37	0,60	0,60	0,60	0,34	0,57	0,51	2,56
njw	0,62	0,59	0,62	0,65	0,62	0,64	0,61	0,54	0,60	1,55

Ghi chú: kch: không chuẩn hóa; sm: chuẩn hóa theo Shi and Malik (2000); njw: chuẩn hóa theo Ng et al. (2002); d1 - d9: elu1, elu2, elu3, cdc25 - 1, cdc25 - 2.1, cdc25 - 2.2, cdc25 - sep1, elu - cdc10, và elu - cdc25; R^a: bậc trung bình.

(chỉ số Rand < 0,5). Thuật toán phân cụm phổ chuẩn hóa theo Ng et al. (2002) có hiệu năng tốt nhất. Kiểm định Friedman phát hiện có khác biệt giữa các thuật toán $X^2 = 15{,}3, df = 2, p = 0{,}01$.

4. KẾT LUẬN

Trong bài báo này, ba thuật toán phân cụm phổ phổ biến nhất đã được nghiên cứu: phân cụm phổ không chuẩn hóa, phân cụm phổ chuẩn hóa theo Shi and Malik (2000), phân cụm phổ chuẩn hóa theo Ng et al. (2002). Khoảng cách xoắn thời gian động được áp dụng để đo độ tương tự giữa các hồ sơ biểu diễn gene. Bốn độ đo hiệu lực được sử dụng để đánh giá chất lượng, độ ổn định, và chính xác của phân cụm. Thuật toán phân cụm phổ chuẩn hóa theo Ng et al. (2002) được chỉ ra là tốt hơn so với hai thuật toán phổ còn lại dưới hai độ đo hiệu lực phân cụm là chỉ số Silhouette và chỉ số Rand. Như những kết quả đánh giá được đưa ra, thuật toán phân cụm phổ chuẩn hóa theo Ng et al. (2002) nên được tiến cử trong các thuật toán phân cụm phổ cho ứng dụng phân cụm dữ liệu biểu diễn gene.

TÀI LIỆU THAM KHẢO

Al - Naymat G., S. Chawla, and J. Teheri (2009). SparseDTW: a novel approach to speed up dynamic time warping. Proc. of the eighth Australasian data mining conference, 101: 117 - 127.

Bayá E. A., and P. M. Granitto (2011). Clustering gene expression data with a penalized graph - based metric. BMC Bioinformatics, 12: 2.

Boeva V., and E. Tsiporkova (2010). A multi - purpose Time series data standardization method. Intelligent systems: from theory to practice, Springer - Varlag Berlin Heidelberg, 299: 445 - 460.

Borg A., N. Lavesson, and V. Boeva (2013). Comparison of clustering approaches for gene expression data. Twelfth scandinavian conference on artificial intelligence. Jaeger M. (ed), 2013.

Chen Y., M. Dong, and M. Rege (2007). Gene expression clustering: a novel graph partitioning

approach. International joint conference on neural networks, p. 1542 - 1547.

Chung F. (1997). Spectral graph theory. The CBMS regional conference series in mathematics, Washington.

Datta S. (2003). Comparisons and validation of statistical clustering techniques for microarray gene expression data. Bioinformatics, 19(4): 459 - 466.

Frey B. J., and D. Dueck (2007). Clustering by passing messages between data points. Science, 5814: 972 - 976.

Huang G. T., K. I. Cunningham, P. V. Benos, and C. S. Chennubhotla (2013). Spectral clustering strategies for heterogeneous disease expression data. Pac Symp Biocomput, 2013: 212 - 223.

Handl J., J. Knowles, and D. B. Kell (2005). Computational cluster validation in post - genomic data analysis. Bioinformatics, 21: 3201 - 3212.

Jaccard P. (1912). The distribution of flora in the alpine zone. New Phytologist, 11: 37 - 50.

Jiang D., C. Tang, and A. Zhang (2004). Cluster analysis for gene expression data: a survey. IEEE Transactions on Knowledge and Data Engineering, 16(11): 1370 - 1386.

Luxburg V. U. (2007). A tutorial on spectral clustering. Stat Comput, 17(4): 395 - 416.

Mohar B. (1991). The Laplacian spectrum of graphs. *In:* Graph theory, combinatorics, and applications, 2, Kalamazoo, MI (1988), p. 871 - 898, New York, Wiley.

Mohar B. (1997). Some applications of Laplace eigenvalues of graphs. *In:* Graph Symmetry: Algebraic Methods and Applications, Hahn G. and G. Sabidussi (Eds.), NATO ASI Ser. C 497: 225 - 275, Kluwer.

Ng, A., M. Jordan, and Y. Weiss (2002). On spectral clustering: analysis and an algorithm. *In:* Advances in Neural Information Processing Systems, Dietterich T., S. Becker, and Z. Ghahramani (Eds.), MIT Press, 14: 849 - 856.

Nguyen V. A., and P. Li (2009). Measuring similarity between gene expression profiles: a Bayesian approach. BMC Genomics, 10(3): 1 - 10.

Tsiporkova E., and V. Boeva (2007). Two - pass imputation algorithm for missing value estimation in gene expression time series. Journal of Bioinformatics and Computational Biology, 5(5): 1005 - 1022.

Rousseeuw P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational applied mathematics, 20: 53 - 65.

Rustici G., J. Mata, K. Kivinen, P. Lió, C. J. Penkett, G. Burns, J. Hayles, A. Brazma, P. Nurse and J. Bähler (2004). Periodic gene expression program of the fission yeast cell cycle. Nature genetics, 36(8): 809 - 817.

Quackenbush J. (2001). Computational analysis of microarray data. Nature Reviews Genetics, 2(6): 418 - 427.

Shi, J., and J. Malik (2000). Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8): 888 - 905.

Xu Y., V. Olman, and D. Xu (2002). Clustering gene expression data using a graph - theoretic approach: an application of minimum spanning trees. Bioinformatics, 18(4): 536 - 545.

Rand W. M. (1971). Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 66: 846 - 850.