SPARSE STAR COORDINATES: VISUALIZATION FOR HIGH DIMENSION LOW SAMPLE SIZE

Tran Van Long*, Bui Viet Huong
University of Transport and Communications, Hanoi

ARTICLE INFO

Received: 17/4/2023
Revised: 24/5/2023
Published: 24/5/2023

KEYWORDS

Star coordinates
High dimension low sample size
Data visualization
Silhouette coefficient
Feature Importance

ABSTRACT

The visual analysis of group structures and trends of high-dimensional data is a central topic in many fields, particularly in genomic data analysis. Gene expression data have a small number of observations and a large number of attributes. The traditional statistical methods are not directly applied to analyze for high dimension, low sample size. In this paper, we introduce a new visualization technique approach to visual analytics of high-dimension, low-sample size. We propose a sparse star coordinates visualization technique based on star coordinates that group structures are preserved thanks to the optimal layouts of star coordinate systems on the visual space. The larger star coordinates are more important dimensions in cluster analysis. The sparse star coordinate system attains by ranking the best quality visualization of the order of the dominant attributes to analyze the group structures of the high-dimension, low-sample size data sets. We present our proposed method with quality measurement and attest to the effectiveness of our approach for several real data sets.

HỆ TOẠ ĐỘ HÌNH SAO THƯA: TRỰC QUAN HÓA DỮ LIỆU SỐ CHIỀU LỚN CỠ MẪU NHỎ

Trần Văn Long*, Bùi Việt Hương
Trường Đại học Giao thông vận tải, Hà Nội

THÔNG TIN BÀI BÁO

Ngày nhận bài: 17/4/2023
Ngày hoàn thiện: 24/5/2023
Ngày đăng: 24/5/2023

TỪ KHÓA

Hệ tọa độ hình sao
Số chiều lớn cỡ mẫu nhỏ
Trực quan hóa dữ liệu
Hệ số Silhouette
Thuộc tính quan trọng

TÓM TẮT

Phân tích khai phá về các cấu trúc nhóm và xu hướng của dữ liệu nhiều chiều là chủ đề chính của nhiều lĩnh vực nghiên cứu có nhiều ứng dụng, đặc biệt trong phân tích dữ liệu gen. Dữ liệu gen có số chiều lớn và số quan sát nhỏ. Các phương pháp phân tích thống kê truyền thống thông thường không được áp dụng trực tiếp cho dữ liệu có số chiều cao, số mẫu nhỏ. Trong bài báo này, chúng tôi giới thiệu cách tiếp cận phân tích dữ liệu bằng trực quan hoá đối với dữ liệu có số chiều cao và cỡ mẫu nhỏ. Chúng tôi đề xuất phương pháp chiếu thưa dựa vào phương pháp trực quan hoá bằng hệ toạ độ hình sao mà cấu trúc nhóm được bảo toàn nhờ vào việc tối ưu hoá sự phân bố hệ toạ độ hình sao. Phương pháp chiếu thưa nhận được từ việc xếp hạng chất lượng trực quan hoá theo thứ tự các thuộc tính quan trọng để lựa chọn các thuộc tính quan trọng trong phân tích cấu trúc nhóm của dữ liệu. Các kết quả thực nghiệm chứng tỏ sự hiệu quả của phương pháp đề xuất.

DOI: https://doi.org/10.34238/tnu-jst.7768

* Corresponding author. *Email: vtran@utc.edu.vn*

1. Giới thiệu

Trực quan hoá phân tích dữ liệu là phương pháp để khám phá về cấu trúc, xu hướng, mối liên hệ giữa các nhóm, mối liên hệ giữa các thuộc tính được sử dụng trong phân tích dữ liệu nhiều chiều. Việc hiểu được sự ảnh hưởng của các thuộc tính đối với một cấu trúc nào đó của dữ liệu rất quan trọng trong phân tích dữ liệu. Các phương pháp biểu diễn dữ liệu nhằm giảm số chiều của dữ liệu bằng các phương pháp chiếu phi tuyến thông thường sẽ bảo toàn một số cấu trúc nào đó của dữ liệu. Tuy nhiên, các phương pháp này không có sự tác động của các thuộc tính trong biểu diễn.

Các phương pháp biểu diễn trực quan hoá dữ liệu nhiều chiều có sử dụng trực tiếp các thuộc tính đối với dữ liệu như phương pháp ma trận biểu đồ phân tán (Scatterplot Matrix – biểu diễn tất cả các cặp thuộc tính), Hệ toạ độ song song (Parallel Coordinates – biểu diễn các điểm bằng các đường gấp khúc), Hệ toạ độ hình sao (Star Coordinate-biểu diễn bằng tổ hợp tuyến tính của hệ véc-tơ trong không gian hai chiều), Hệ toạ độ hướng tâm (Radviz -biểu diễn bởi điểm cân bằng trong hệ lò xo) được giới thiệu trong tổng quan về trực quan hoá [1]. Các phương pháp biểu diễn dữ liệu thường áp dụng đối với các dữ liệu có số chiều cỡ trung bình (dưới 50 chiều). Đối với số chiều lớn thì do hạn chế của sự biểu diễn hoặc có nhiều số chiều nhiễu nên sự biểu diễn không bảo toàn được các cấu trúc của dữ liệu.

Trực quan hoá dữ liệu nhiều chiều để hiểu cấu trúc của dữ liệu, để hiểu và khai phá dữ liệu. Các nhà nghiên cứu đã giới thiệu nhiều phương pháp để biểu diễn dữ liệu nhiều chiều và được tổng kết trong bài báo [1]. Trong phần này chúng tôi tiếp cận phương pháp giảm số chiều trong biểu diễn dữ liệu nhiều chiều, dữ liệu biểu diễn bởi các điểm trong không gian trực quan hoá (2D). Chúng tôi tổng kết một số nghiên cứu gần đây về phương pháp hệ toạ độ hình sao và phương pháp Radviz.

Hệ toạ độ hình sao (Star Coordinates) là phương pháp biểu diễn tuyến tính chiếu dữ liệu nhiều chiều xuống không gian biểu diễn hai chiều được Kandogan [2] giới thiệu đầu tiên với các trục phân phối đều trên đường tròn đơn vị. Bài báo tiếp theo của cùng tác giả [3] giới thiệu về các phương pháp tương tác với hệ toạ độ hình sao trong biểu diễn dữ liệu nhiều chiều.

Những năm gần đây nhiều tác giả đã nghiên cứu về phương pháp biểu diễn hệ toạ độ hình sao và các phương pháp tương tác. Trong [4], các tác giả nghiên cứu phương pháp tương tác về nhóm các số chiều đối với dữ liệu có số chiều lớn trong biểu diễn dữ liệu. Wang và các cộng sự giới thiệu về phương pháp tối ưu hoá biểu diễn hệ toạ độ hình sao [5] trong bảo toàn cấu trúc nhóm của dữ liệu và sử dụng hệ số *silhouette* để đánh giá chất lượng của biểu diễn trực quan. Rave và các cộng sự [6] đề xuất phương pháp gộp các số chiều trong biểu diễn và tương tác với hệ toạ độ hình sao trong biểu diễn dữ liệu. Năm 2018, Sanchez và các cộng sự [7] nghiên cứu về ứng dụng của hệ tọa độ hình sao trong phân tích dữ liệu y học. Trong bài báo này, các tác giả đã sử dụng độ lớn của hệ toạ hình sao tương ứng với việc khôi phục lại dữ liệu nghĩa là biểu diễn bảo toàn cấu trúc dữ liệu ban đầu. Năm 2021, Alberto Sanchez và các cộng sự [8] ứng dụng phương pháp hệ toạ độ hình sao để đánh giá các thuộc tính quan trọng dựa vào việc tính các giá trị riêng bằng cách đưa ra đánh giá độ quan trọng của các thuộc tính tương ứng với độ lớn của các trục biểu diễn các thuộc tính trong hệ toạ độ hình sao.

Hệ toạ độ hướng tâm (Phương pháp Radviz) là phương pháp biểu diễn dữ liệu tương tự như phương pháp hệ toạ độ hình sao. Đây là phương pháp chiếu phi tuyến và được giới thiệu đầu tiên bởi Hoffman và các cộng sự [9]. Rubio-Sánchez và các cộng sự [10] nghiên cứu về mối quan hệ giữa hai phương pháp biểu diễn Radviz và hệ toạ độ hình sao. Các đề xuất cải tiến của phương pháp Radviz được các tác giả nghiên cứu gần đây như phương pháp VizRank [11], phương pháp FreeViz [12], PolarViz [13], ArcViz [14]. Phương pháp RadViz được ứng dụng trong biểu diễn dữ liệu gen và phân loại các loại gen có ảnh hưởng đến một số loại ung thư thông qua phương pháp biểu diễn bảo toàn cấu trúc nhóm có bệnh và nhóm không có bệnh được các tác giả công bố trong công trình [15].

Trong phân tích dữ liệu y học đặc biệt về dữ liệu gen đối với một số bệnh ung thư nào đó chúng ta cần xác định được nhóm các gen có tác động ảnh hưởng đến bệnh. Đối với dữ liệu gen số quan sát là số bệnh nhân (n) cỡ vài trăm và số thuộc tính là số các gen (p) cỡ mười nghìn. Đây là kiểu dữ liệu có số chiều lớn và số quan sát nhỏ. Với các phương pháp phân tích thống kê truyền thống thường chỉ áp dụng được đối với các dữ liệu có số quan sát lớn và số chiều nhỏ hơn số quan sát, còn với dữ liệu có $n \ll p$ có nhiều thuộc tính không có ảnh hưởng nhiều đến cấu trúc dữ liệu, nghĩa là chỉ có một số các thuộc tính có ảnh hưởng đến cấu trúc dữ liệu. Các phương pháp được sử dụng trong phân tích đối với dữ liệu có số chiều cao, cỡ mẫu nhỏ gồm có phương pháp giảm số chiều tuyến tính phân tích thành phần chính (PCA) và phương pháp phân tích thành phần phân biệt (LDA).

Trong bài báo này chúng tôi đề xuất phương pháp trực quan hoá để phân tích dữ liệu có số quan nhỏ và số chiều lớn. Chúng tôi sử dụng phép chiếu lên hệ toạ độ hình sao đối với toàn bộ các thuộc tính và đánh giá chất lượng của các diễu diễn thông qua mạng trí tuệ nhân tạo để phân loại theo các nhóm và đồng thời đánh giá các thuộc tính quan trọng trong biểu diễn của hệ toạ độ hình sao. Để xác định chất lượng biểu diễn hiệu quả đối với cấu trúc nhóm chúng ta lựa chọn một số các thuộc tính quan trọng trong hệ toạ độ hình sao dựa vào chất lượng biểu diễn trực quan thông qua đánh giá hệ số *silhouette*. Chúng tôi đề xuất phương pháp chiếu thưa trong biểu diễn dữ liệu có $n \ll p$. Kết quả thực nghiệm đối với 8 dữ liệu gen và các phân tích với kết quả thu được cho thấy phương pháp đề xuất là hữu hiệu.

2. Phương pháp nghiên cứu

2.1. Hệ tọa độ hình sao

Phương pháp chiếu bằng hệ toạ độ hình sao từ không gian nhiều chiều xuống không gian trực quan hoá là phương pháp biến đổi tuyến tính. Trong phương pháp hệ toạ độ hình sao mỗi thuộc tính của dữ liệu được biểu diễn bởi một véc-tơ hai chiều và điểm biểu diễn dữ liệu nhiều chiều là tổ hợp tuyến tính của các thuộc tính với trọng số là giá trị của các thành phần của dữ liệu. Cụ thể, cho điểm trong không gian dữ liệu nhiều chiều $x = (x_1, x_2, \dots, x_p)$ với hệ toạ độ hình sao $V = (V_1, V_2, \dots, V_p)$ trong đó mỗi véc-tơ $V_i = (V_{i1}, V_{i2})$ biểu diễn số chiều thứ i. Phép chiếu bằng hệ toạ độ hình sao biểu diễn dữ liệu x bởi điểm y trong không gian biểu diễn xác định bởi công thức:

$$y = \sum_{i=1}^{p} x_i V_i \ . \quad (1)$$

Trong các bài báo [2], [3], hệ toạ độ hình sao được xác định bởi các véc-tơ V_i biểu diễn đều trên đường tròn đơn vị với $V_i = \left(\cos\frac{2\pi(i-1)}{p}, \sin\frac{2\pi(i-1)}{p}\right), \ i = 1, \dots, p.$

2.2. Phương pháp chiếu thưa

Cho dữ liệu $X = (x_{ij})_{n\times p}$ gồm n quan sát trong không gian p chiều và được phân loại thành K lớp. Ký hiệu ma trận $y = (y_{ij})_{n\times K}$, trong đó $y_{ij} = 1$ nếu dữ liệu thứ i thuộc lớp thứ j và $y_{ij} = 0$ nếu trái lại. Trong bài báo này, chúng ta nghiên cứu bài toán tìm phép chiếu bằng hệ toạ độ hình sao để bảo toàn sự phân lớp của dữ liệu từ không gian dữ liệu xuống không gian trực quan hoá.

Đối với dữ liệu có số chiều lớn và số quan sát nhỏ ($n \ll p$) thì không gian biểu diễn dữ liệu sẽ được phân loại hoàn hảo trong một không gian có số chiều bé hơn không gian dữ liệu. Trong phần này, chúng tôi đề xuất phương pháp chiếu thưa bằng hệ toạ độ hình sao để bảo toàn dữ liệu phân lớp của dữ liệu trong không gian biểu diễn dữ liệu.

Để tối ưu hóa hệ tọa độ hình sao, chúng tôi đề xuất sử dụng mạng trí tuệ nhân tạo. Xét hệ toạ độ hình sao $V = (v_{ij})_{2\times p}$ để chiếu dữ liệu từ không gian dữ liệu nhiều chiều xuống không gian biểu

diễn dữ liệu hai chiều bằng công thức (1). Để phân loại dữ liệu trong không gian biểu diễn, chúng tôi sử dụng phương pháp hồi quy logistic với hàm kích hoạt relu được xác định theo công thức:

$$\text{relu}(x) = \max\{0, x\}$$

và hàm softmax xác định bởi

$$\text{softmax}(z) = \frac{1}{\sum_{i=1}^{n} \exp(z_i)} (\exp(z_1), \exp(z_2), \dots, \exp(z_n)).$$

Việc xác định tối ưu hoá sự phân loại, chúng tôi dựa trên công thức tính xác suất phân loại cho điểm dữ liệu x_i như sau

$$a_i = s\text{oftmax}\,(W\ \text{relu}(Vx_i) + b), \quad (2)$$

trong đó $W = \left(w_{ij}\right)_{2\times K}$ và véc-tơ $b = (b_1, \dots, b_K)$. Hàm thất thoát được xác định bằng phương pháp cực tiểu hoá entropy chéo giữa xác suất phân loại a_i và sự phân lớp y_i cho dữ liệu thứ i, và hàm thất thoát cho toàn bộ dữ liệu xác định bởi công thức sau:

$$L\,(V, W, b) = \sum_{i=1}^{n} \left(-\sum_{j=1}^{n} y_{ij} \log(a_{ij}) \right). \quad (3)$$

Chúng ta cần xác định một số các thuộc tính quan trọng tương ứng với các véc-tơ V_i có độ dài lớn nhất trong biểu diễn dữ liệu. Khi đó hàm thất thoát được cộng thêm với trọng số xác định độ lớn của hệ toạ độ hình sao bằng chuẩn trong không gian L_1 với $\|V\|_1 = \sum_{ij} |v_{ij}|$.

Hàm tối ưu đối với hệ toạ độ hình sao thưa được xác định bởi công thức

$$J(V, W, b) = L(V, W, b) + \lambda \|V\|_1, \quad (4)$$

với λ là trọng số dương.

2.3. Chất lượng trực quan hóa

Để xác định chất lượng của trực quan hoá đối với dữ liệu phân loại, chúng tôi đề xuất sử dụng hệ số *silhouette* [5]. Hệ số *silhouette* được sử dụng để đánh giá kết quả của sự phân lớp, hệ số này nằm trong khoảng $[-1,1]$ và hệ số càng lớn thì hiệu quả của sự phân lớp càng cao.

Để tối ưu hoá hệ toạ độ hình sao ta sắp xếp hệ tọa độ hình sao theo độ dài của các véc-tơ biểu diễn $||V_i||, i = 1,2, \dots, p$. Hệ toạ độ hình sao thưa là hệ toạ độ hình sao bao gồm q thuộc tính có độ dài lớn nhất có chất lượng biểu diễn bằng hệ số *silhouette* cao nhất có thể.

3. Thực nghiệm và kết quả

Trong phần này chúng tôi trình bày một số kết quả thực nghiệm đối với dữ liệu thực tế, cụ thể là dữ liệu gen. Đối với dữ liệu gen thì số quan sát là số các bệnh nhân (cỡ khoảng 100 mẫu) và số thuộc tính là số gen (cỡ khoảng 10000 thuộc tính). Đối với dữ liệu gen, chúng ta cần xác định nhóm các gen có ảnh hưởng đến một số bệnh nào đó.

3.1. Dữ liệu

Bảng 1. *Bảng mô tả dữ liệu gen*

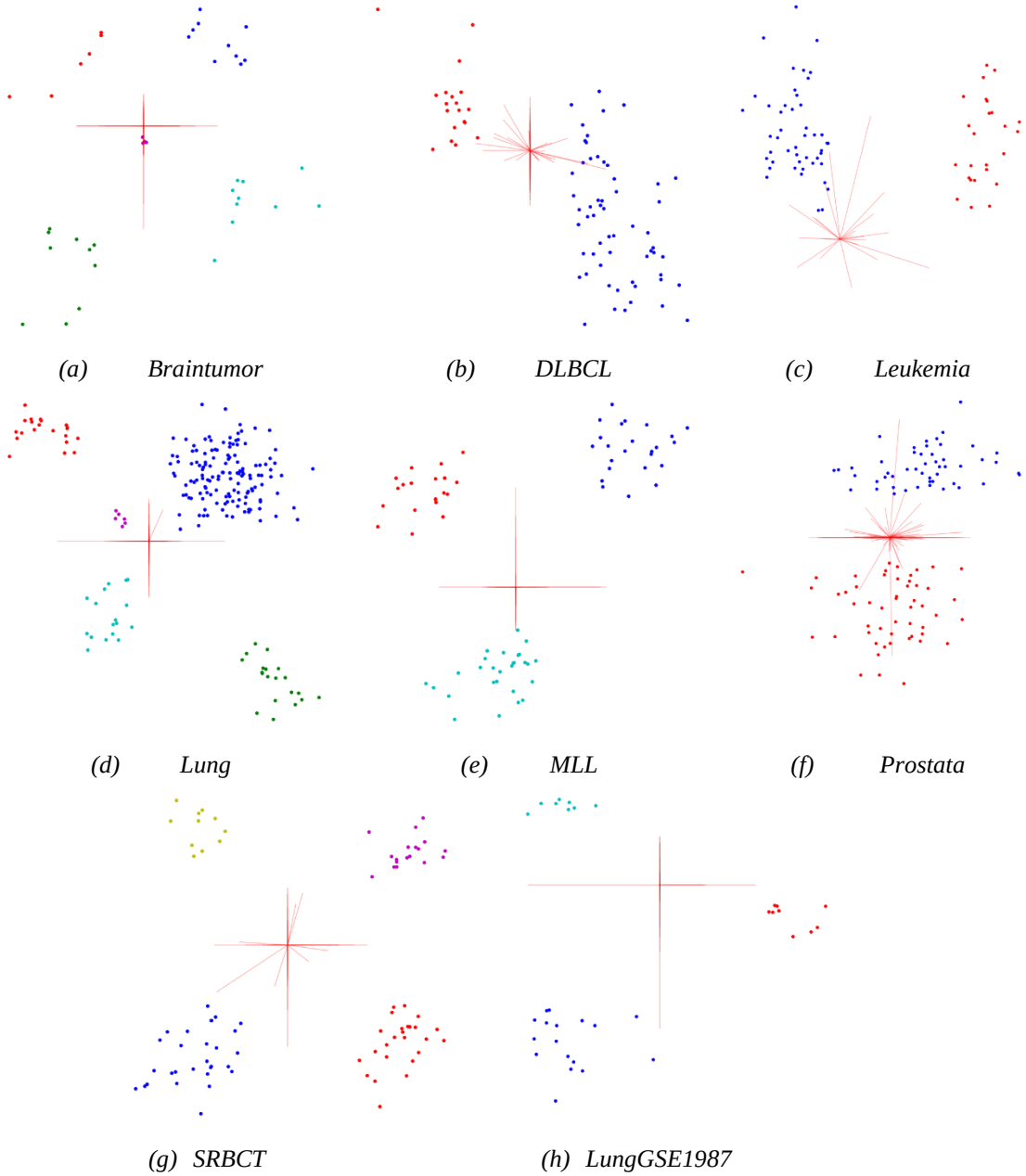
Dữ liệu	Số mẫu	Số chiều	Số lớp
Braintumor	40	7129	5
DLBCL	77	7070	2
Leukemia	72	5147	2
Lung	203	12600	5
LungGSE1987	34	10541	3
MLL	72	125333	3
Prostata	102	125333	2
SRBCT	83	2308	4

Chúng tôi sử dụng 8 bộ dữ liệu gen (*https://file.biolab.si/biolab/supp/bi-cancer/projections/*) được mô tả trong Bảng 1. Trong đó số quan sát từ 34 đến 203, số thuộc tính từ 2308 đến 12600,

và số lớp phân loại từ 2 đến 5 lớp. Các dữ liệu trên đều là các dữ liệu có số quan sát n nhỏ và số chiều p lớn.

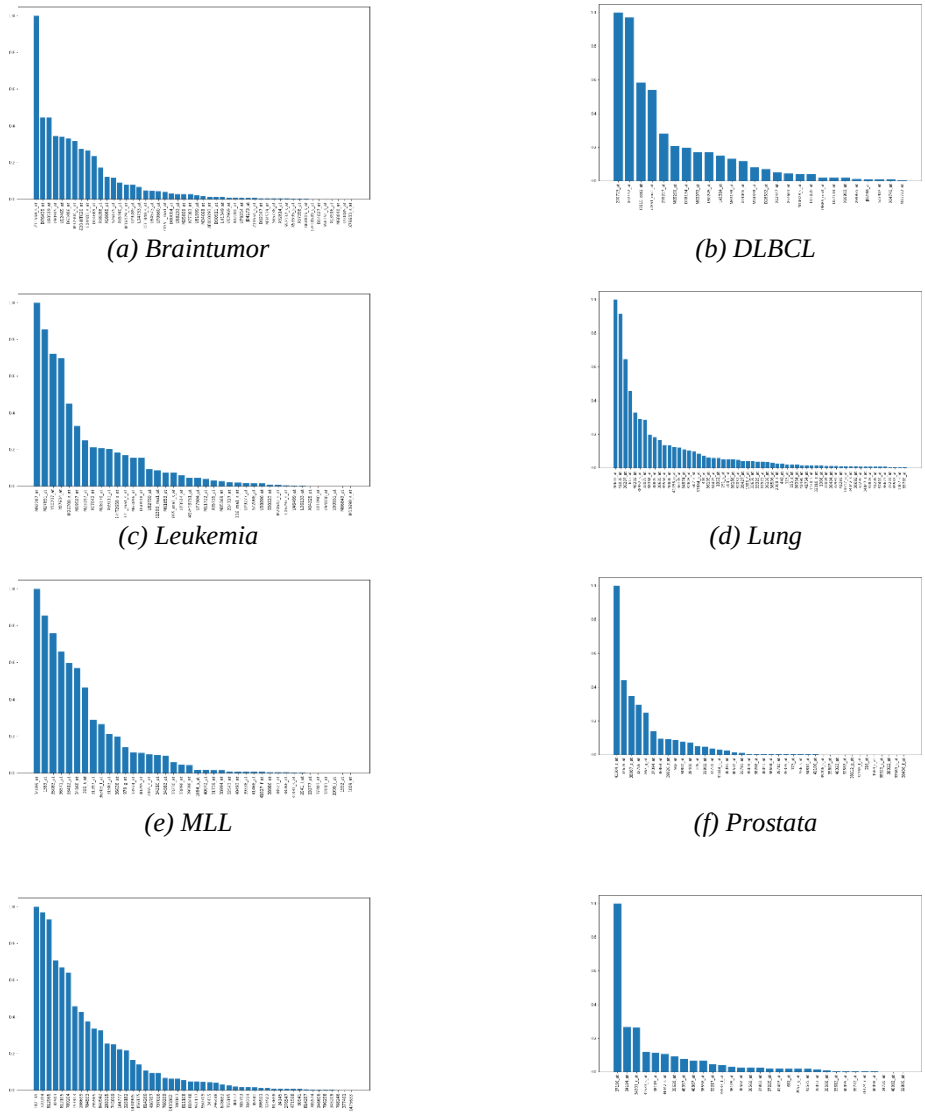
3.2. Tham số thực nghiệm

Trong toàn bộ các kết quả thực nghiệm, để tối ưu hoá hàm mục tiêu (4), chúng tôi sử dụng các tham số $\lambda = 100$, tốc độ học $learning\ rate = 0.001$, tốc độ giảm $decay\ rate = 0.95$, số bước giảm tốc độ học $step\ size = 1000$, và số bước lặp $number\ of\ epochs = 50000$. Các kết quả số được mô tả trong hình 1 và 2.



Hình 1. *Tối ưu hoá hệ toạ độ hình sao thưa đối với dữ liệu (a) Braintumor, (b) DLBCL, (c) Leukemia, (d) Lung, (e) MLL, (f) Prostata, (g) SRBCT và (h) LungGSE1987*

Hình 1 trình bày kết quả về trực quan hóa biểu diễn các dữ liệu gen. Các kết quả thể hiện biểu đồ phù hợp với độ lớn của hệ toạ độ hình sao tương ứng với số thuộc tính lớn nhất và đó cũng là các thuộc tính quan trọng của dữ liệu trong biểu diễn phân loại các nhóm dữ liệu được trình bày trong Hình 2.



(a) Braintumor *(b) DLBCL*

(c) Leukemia *(d) Lung*

(e) MLL *(f) Prostata*

(g) SRBCT *(i) LungGSE1987*

Hình 2. *Các thuộc tính quan trọng đối với dữ liệu (a) Braintumor, (b) DLBCL, (c) Leukemia, (d) Lung, (e) MLL, (f) Prostata, (g) SRBCT và (h) LungGSE1987*

3.3. Kết quả thực nghiệm

Bảng 2 trình bày các kết quả thực nghiệm về các dữ liệu gen bao gồm chất lượng biểu diễn trực quan đối với hệ số *silhouette* và số thuộc tính có chất lượng biểu diễn trực quan tốt nhất bằng hệ toạ độ hình sao. Ở đây, chúng tôi so sánh kết quả với hai phương pháp biểu diễn dữ liệu tuyến tính phổ biến là phương pháp phân tích thành phần chính (PCA) và phương pháp phân tích thành phần phân biệt (LDA) (xem [15]). Kết quả cho thấy, đối với hầu hết dữ liệu, phương pháp chiếu

thưa đạt kết quả cao hơn so với các phương pháp khác ngoại trừ dữ liệu SRBCT được biểu diễn bằng phương pháp LDA.

Bảng 2. *Kết quả thực nghiệm*

Dữ liệu	Silhouette hình sao thưa	Số tọa độ hình sao thưa	Silhouette PCA	Silhouette LDA
Braintumor	**0.7994**	39	0.1477	0.1189
DLBCL	**0.8123**	76	0.10366	0.5852
Leukemia	**0.8699**	22	0.3029	0.5219
Lung	**0.6117**	190	0.1312	0.5863
LungGSE1987	**0.8299**	33	0.0913	0.2343
MLL	**0.8456**	71	0.2832	0.4237
Prostata	**0.7638**	100	0.0777	0.6268
SRBCT	0.7909	82	-0.0928	**0.9118**

4. Kết luận

Bài báo trình bày về phương pháp chiếu thưa dựa vào hệ toạ độ hình sao thông qua phương pháp tối ưu hoá của mạng trí tuệ nhân tạo và hệ toạ độ hình sao thưa được đánh giá thông qua độ lớn tương ứng với độ quan trọng của các thuộc tính. Các kết quả thực nghiệm đánh giá phương pháp chiếu thưa cho kết quả tốt đối với hầu hết các dữ liệu, các nhóm dữ liệu được tách nhau khá hoàn hảo trong không gian biểu diễn. Trong phần nghiên cứu tiếp theo chúng tôi sẽ nghiên cứu về phương pháp đánh giá mức độ quan trọng của các thuộc tính biểu diễn đối với phương pháp biểu diễn Radviz đối với dữ liệu có số quan sát nhỏ và số chiều lớn.

Lời cảm ơn

Nghiên cứu này được tài trợ bởi Trường Đại học Giao thông vận tải trong đề tài mã số T2023 – CB – 010.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] L. Shusen, M. Dan, W. Bei, P. Bremer, and V. Pascucci, "Visualizing high-dimensional data: Advances in the past decade," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 3, pp. 1249-1268, 2017.

[2] E. Kandogan, "Star coordinates: A multi-dimensional visualization technique with uniform treatment of dimensions," *Proceedings of the IEEE Information Visualization Symposium, Hot Topics*, 2000, pp. 4-8.

[3] E. Kandogan, "Visualizing multi-dimensional clusters, trends, and outliers using star coordinates," *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD' 01*, 2001, pp. 107-116.

[4] G. Z. Germain, G. N. Luis, and G. Erick, "iStar (i*): An interactive star coordinates approach for high-dimensional data exploration," *Computers and Graphics*, vol. 60, pp. 107-118, 2016.

[5] W. Yunhai, L. Jingting, N. Feiping, T. Holger, G. Minglun, and J. L. Dirk, "Linear Discriminative Star Coordinates for Exploring Class and Cluster Separation of High Dimensional Data," *Computer Graphics Forum*, vol. 36, no. 3, pp. 401-410, 2017.

[6] H. Rave, V. Molchanov, and L. Linsen, "Axes Bundling and Brushing in Sta Coordinates," *International Symposium on Vision, Modeling, and Visualization*, 2021, doi: 10.2312/vmv.20211365.

[7] A. Sanchez, C. Soguero-Ruiz, I. Mora-Jiménez, F. J. Rivas-Flores, D. J. Lehmann, and M. Rubio-Sánchez, "Scaled radial axes for interactive visual feature selection: A case study for analyzing chronic conditions," *Expert Systems with Applications*, vol. 100, pp. 182-196, 2018.

[8] A. Sanchez, L. Raya, M. A. Mohedano-Munoz, and M. Rubio-Sánchez, "Feature selection based on star coordinates plots associated with eigenvalue problems," *The Visual Computer*, vol. 37, pp. 203–216, 2021.

[9] P. Hoffman, G. Grinstein, K. Marx, I. Grosse, and E. Stanley, "DNA visual and analytic data mining," *Proceedings of the 8th conference on Visualization'97*, 1997, pp. 437-441.

[10] M. Rubio-Sánchez, L. Raya, F. Díaz, and A. Sanche, "A comparative study between RadViz and Star Coordinates," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 1, pp. 619-628, 2016.

[11] G. Leban, B. Zupan, G. Vidmar, and I. Bratko, "VizRank: Data visualization guided by machine learning," *Data Mining and Knowledge Discovery*, vol. 13, no. 2, pp. 119-136, 2006.

[12] J. Demsar, G. Leban, and B. Zupan, "FreeViz: An intelligent multivariate visualization approach to explorative analysis of biomedical data," *Journal of Biomedical Informatics*, vol. 40, no. 6, pp. 661-671, 2007.

[13] Y. C. Wang, Q. Zhang, F. Lin, C. K. Goh, and H. S. Seah, "PolarViz: A discriminating visualization and visual analytics tool for high-dimensional data," *The Visual Computer*, vol. 35, pp. 1567–1582, 2019.

[14] T. V. Long, "ArcViz: An Extended Radial Visualization for Classes Separation of High Dimensional Data," *The 10th International Conference on Knowledge and Systems Engineering (KSE 2018)*, 2018, pp. 158-162.

[15] J. F. McCarthy, K. Marx, P. E. Hoffman, A. G. Gee, P. O'Neil, M. Ujwal, and J. Hotchkiss, "Applications of Machine Learning and High-Dimensional Visualization in Cancer Detection, Diagnosis and Management," *Annals of the New York Academy of Sciences*, vol. 1020, no. 1, pp. 239 - 262, 2004.