CIRCULARVIZ: AN ENHANCED RADVIZ VISUALIZING FOR MULTIDIMENSIONAL DATA

Van Long Tran[1]

Abstract

Visualizing for multi-dimensional data points in data space is a central topic with many supplications in many branches of science. RadViz visualizes multi-dimensional data points in the $2D$ visual space based on force-based methods. Radviz visualizes information about relationships of multi-dimensional data points of data space. The traditional Radviz method finds the best order of anchor points on the circumference of the circle with radius one to enhance the relationships of multi-dimensional data sets. In our paper, we modify the Radviz method named CircularViz method that improves the Radviz layout for the group's preservation the relationships of multi-dimensional data sets. The idea of the CircularViz method presents data dimensions by circular segments on the circle with radius 1 and finds the most suitable to gaining insights into class structures of multi-dimensional data sets. Our approach make provision an augmentation in visualizing group structures of multi-dimensional data on the $2D$ visual space. The effectiveness of the CircularViz method evidences by quality visualization measurements and several supervised high-dimensional data sets.

Index terms

Data visualization, Radial visualization, high dimensional data, quality visualization.

1. Introduction

Data visualization is used in various science to understand and interpret data. Information visualization techniques display graphically representing abstract data which supports interactive methods to exploring data structures. Visualization data set with high dimension is a dimensionality reduction method that maps data points in the data space into a visual space (2D or 3D) to maintain relationships of the original data.

Star coordinates introduced in [1], [2] that is point-based technique. Star coordinates maps high-dimensional data into visual space by a linear combination of a set of vectors. Star coordinate provides several interactive techniques for transforming the high-dimensional data set.

The original RadViz [3] and the PolyViz [4] presents multi-dimensional data in the compact form in the 2D visual space. Van Long [5] introduces another variant of the RadViz method named ArcViz for display cluster separation of data points in the data space. The ArcViz has replaced the line segment of attributes with an arc of dimensions.

[1] Faculty of Basic Science, University of Transport and Communications, Hanoi, Vietnam.

The ArcViz supports more area space for the visualization of high-dimensional data in the $2D$ visual space of a circle with radius 1.

In our paper, we introduce the CircularViz method for visualizing data in the data space bayond three dimensions based on the radial visualization technique. The CircularViz has overcome the limitation of the consecutive dimension anchors in the PolyViz and the ArcViz visualization. The CircularViz supports a more efficient space for representing multivariate data inside the circle with radius 1.

Our paper is composed as given follows. In Section 2, some recently point-based technique for multidimensional data visualization and variants of radial visualization method is discribed. Section 3 describes the CircularViz method for projecting data points in the data space into the $2D$ visual space. Section 4 presents the LDC (Linear Discriminant Classifier) method and KNNC (k-Nearest Neighbor Classifiers) to determine quality visualization measurements for objective functions. In Section 5, we validate the usefulness of the CircularViz method over a variety of experimental data sets. The last Section, we present conclusions and future works.

2. Related works

Radial visualization is the first introduced [3] projects high dimensional data into visual space in two-dimensional space. Radviz locates anchor points around the circumference of a circle of radius one. RadViz is a force-based point layout technique that maps data points in data space into the $2D$ display space based on spring systems.

PolyViz [4], [6] is a modified of the RadViz method. PolyViz presents data dimensions on the side of a convex polygon. High-dimensional data point visualizes similar the RadViz method. However, PolyViz shows the distribution of each dimension on the side of the convex polygon.

Concentric RadViz [7], [8] is used to visualize the result of multi-class and multi-level data sets. The Concentric RadViz is used concentric circles for classification tasks and named as Dimension Group. The Concentric RadViz is also improved for reducing clutter in the center of the circle based on the sigmoid weighting technique.

DualRadviz [9] is another modification of the RadViz method that uses double RadViz with the same circle with radius 1, one of them is the traditional Radviz and another RadViz is used to visualizing group structures of the data. The data points in the data space are visualized in the 2D visual space by using the data features and the cluster information.

3. CircularViz Method

For a m-dimensional data set, m anchor points $S_1, S_2, \ldots, S_m$ display on the circumference of the unit. The anchor points $S_1, S_2, \ldots, S_m$ are usually placed on a circle of radius 1 centered on the origin

$$S_k = (\cos\alpha_k, \sin\alpha_k), k = 1, \dots, m, \tag{1}$$

where angles $\alpha_1, \alpha_2, \dots, \alpha_m$ (initial set $\alpha_k = 2\pi(k-1)/m, k = 1, 2, \dots, m$).

In [5] presented the Radial vizsualization, Polygonal visualization, and Arc visualization methods in detail. We advance a new modification of the Radviz method named as CircularViz.

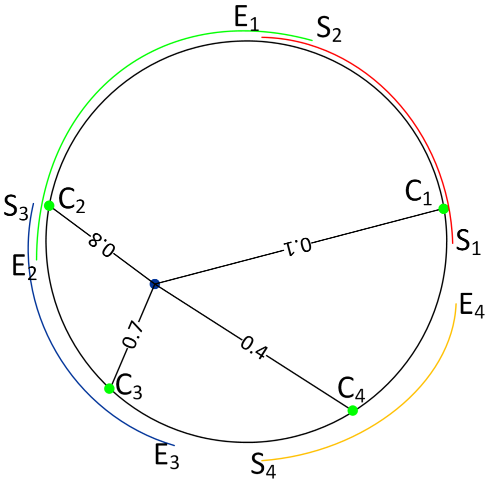


Fig. 1. Four dimensional point display on the $2D$ visual space with CircularViz.

We denote the weights for the m-dimensional point $x = (x_1, x_2, \dots, x_m)$

$$w_k(x) = \frac{x_k}{\sum\limits_{k=1}^{m} x_k}, k = 1, \dots, m. \tag{2}$$

The CircularViz is a new modified of RadViz visualization techniques. The dimensional anchor in the ArcViz describes as a part of the circle unit. The CircularViz is extended of the ArcViz without consecutive an arc on the circle with radius 1. The CircularViz does not depend on the order of attributes of the data sets.

The ith circular anchor is defined as an arc on the circle with radius 1 with ending points S_k and E_k where

$$S_k = (\cos\alpha_k, \sin\alpha_k), E_k = (\cos\beta_k, \sin\beta_k);$$

for each data dimension $k = 1, \dots, m$. A dummy dimensional anchor for a m-dimensional data point $x = (x_1, x_2, \dots, x_m)$ is determined as follows

$$C_k = (\cos\theta_k, \sin\theta_k), k = 1, 2, \dots, m;$$

where $\theta_k = x_k\alpha_k + (1 - x_k)\beta_k, k = 1, \dots, m$. The CircularViz projects observations from a m-dimensional data space to the $2D$ display space. For each record

$x = (x_1, x_2, \ldots, x_m)$ in a hypercube $[0,1]^m$ is represented as a point p in the $2D$ display space by given formula:

$$p = \sum_{k=1}^{m} \omega_k(x) C_k. \tag{3}$$

Figure 1 display the 4-dimensional point on the $2D$ visual space by CircularViz method. In Figure 1, circular anchors jittered to avoid overlapping with the circle with radius 1 and other circular anchors. The CircularViz can view as the RadViz if the circular dimensional anchor degenerates a single point, i.e. $S_k \equiv E_k$, and the CircularViz is the ArcViz if arcs $S_k E_k$ are consecutive on the circle with radius 1, i.e. $S_k \equiv E_{k+1}$.

4. Quality Metrics for Visualization

Given a data set X contains n-observations $(x_1, x_2, \ldots, x_n)$ in a m-dimensional data space. Assume the data set is divided into K classes. We define n_k as the size of the kth group. For a general projecttion P, we denote $y_i = P(x_i)$ and the data set X is visualized by the data set $Y = [y_1, y_2, ..., y_n]$ in the 2D visual space. We define the quality metrics in visual space by using linear classifier and non-linear classifier [10].

4.1. Linear Discriminant Classifiers (LDC)

The linear discriminat functions defined as

$$\ell_k(x) = x^T \hat{\Sigma}^{-1} \mu_k - \frac{1}{2} \mu_k \hat{\Sigma}^{-1} \mu_k + \log \pi_k,$$

where $\pi_k = \frac{n_k}{n}$, μ_k denote the mean of the kth group, and $\hat{\Sigma}$ is the sample covariance matrix of the given data set. The linear discriminant classifier is description of the maximum of the linear discriminant functions

$$G(x) = \arg\max_k \ell_k(x).$$

4.2. K-Nearest-Neighbor Classifiers (KNNC)

The KNN (k-Nearest-Neighbor) classifies by finding the k closest training points in the Euclidean distance to the query point x, and classify using majority vote among the k neighbors.

4.3. Quality Visualization Measurements (QVM)

A m-dimensional data point x_i in the given supervised data set $X = [x_1, x_2, \ldots, x_n]$ is projected at the position y_i in the 2D visual space by the Radviz, PolyViz, ArcViz, and CircularViz methods. The m-dimensional data point x_i is corrected displaying in the 2D visual space if its label $label(y_i)$ (determine by LDC or KNNC) is the same label of x_i, i.e. $label(x_i) = label(y_i)$. The quality metric for visualization is defined by measurement the number of observations that are visualized of their corrected classes, i.e.,

$$QVM = \frac{|\{x_i : label(y_i) = label(x_i)\}|}{n}. \tag{4}$$

5. Experiments and results

5.1. Optimal Visualization

The optimal visualization of the CircularViz is the highest quality measurement of the target function QVM in (4). The objective function $f(\alpha_1, \beta_1, \dots, \alpha_m, \beta_m)$ with respect to the supervised data set X equals QVM. The parameters α_i, β_i is in interval $[0, 2\pi)$ for all $i = 1, \dots, m$.

We apply an evolutionary algorithm that named as differential evolution (DE) algorithm [11] to determine the best angles α_k, β_k of the circular anchors S_kE_k. The candidate solution contains $2m$ parameters in the $[0, 2\pi)$ interval. In the initial step, we create the first population by uniform random $U(0, 2\pi)$ with numbers of population $NP = 75$. For next generation, we use the $DE/rand/1/exp$ strategy [12]. The new candidate is created based on mutation operator with probability $F = 0.47$ and crossover operator with probability $CR = 0.88$. The optimal solution is attained from the candidate solution that corresponds to the highest values of the target function with the number of generation less than 50.

5.2. Results

We demontrate the effectiveness of the CircularViz method and comparison to the Radial visualization, Polygonal visualization, and Arc visualization methods over six experiment data sets, i.e., one synthetic data and five real data sets. The synthetic data that named as $Y14c$ data set. The $Y14c$ data includes 480 observations with 10 features and classifier 14 groups. The Iris data contains 150 observations, 4 features, and 3 groups (Setosa, Vericolor, Virgica). The Wine data includes 178 instances, 13 attributes and 3 groups. The Olive Oil data have 572 data points, 8-dimension, and 9 classes. The Auto MPG data contains 398 observations, 8 attributes, and 3 groups. The Ecoli data includes 336 instances, 8 attributes and 8 groups.

The optimal quality visualization of the CircularViz method for six experimental data sets in Table 1 corresponding to the LDC classifier. We compare the optimal quality visualization of the Radial visualization, Polygonal visualization, and Arc visualization citeLong18. Table 1 shows the CircularViz visualization achieved the quality visualization measurement better than the quality visualization of the Radial visualization, Polygonal visualization, and Arc visualization methods.

Table 2 shows the quality visualization measurement of the CircularViz method based on the KNNC of visualization for testing data sets. The quality visualization

Table 1. Results of Radviz, PolyViz, ArcViz and CircularViz based on LDC.

Data sets	Original	RadViz	PolyViz	ArcViz	CircularViz
Auto-mpg	76.27%	75.51%	73.21%	73.47%	**76.02%**
Ecoli	88.69%	74.70%	81.84%	77.98%	**84.22%**
Iris	98.00%	84.67%	97.33%	98.00%	**99.33%**
Olive	94.76%	**90.20%**	84.61%	82.34%	89.69%
Wine	98.88%	**95.51%**	93.25%	92.13%	**95.51%**
Y14c	100%	93.75%	93.13%	98.75%	**100%**

of the CircularViz method shows higher the quality of the original data sets. We also tested the best quality visualization measurement of the Radial visualization, Polygonal visualization, and Arc visualization method and show in Table 2. The class separation of testing data sets based on the KNNC classifier of the CircularViz method is higher than the class separation of the Radial visualization, Polygonal visualization, and Arc visualization.

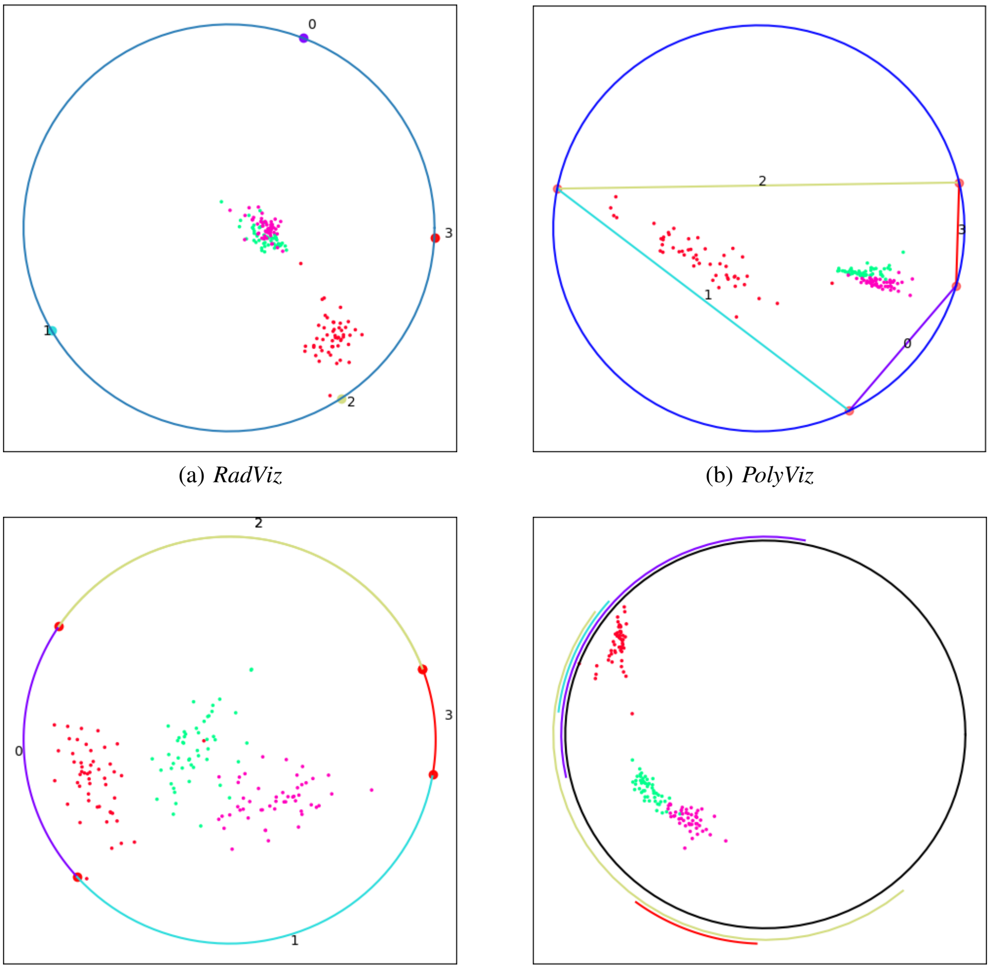
Table 2. Results of Radviz, PolyViz, ArcViz and CircularViz based on KNNC.

Data sets	Original	RadViz	PolyViz	ArcViz	CircularViz
Auto-mpg	71.67%	**80.61%**	78.06%	77.55%	80.36%
Ecoli	87.20%	80.36%	81.84%	82.74%	**85.12%**
Iris	98.00%	89.33%	97.33%	98.00%	**99.33%**
Olive	94.96%	89.69%	87.06%	85.66%	**90.21%**
Wine	81.46%	**96.07%**	94.94%	89.89%	**96.07%**
Y14c	100%	96.88%	95.21%	99.38%	**100%**

The Iris data have 150 observations in 4-dimensional data space and classification into three groups. The Iris gropus data encoded in different colors. Figure 2 shows the optimal visualization of the RadViz, PolyViz, ArcViz, and CircularViz method based on the LDC classifier. Figure 2d displays the largest classification visualization for groups separation of this data set. The CircularViz supports more space for representation the Radial visualization, Polygonal visualization, and Arc visualization, but The CircularViz presents clusters much more compactness than the ArcViz visualization.

The Y14c data contains 14 groups in ten-dimensional space and perfectly separation from the LDC classifier in the original data set. Clusters of data set are encoded in different colors. The optimal visualization of Radviz and Polyviz shows in Figure 3a and Figure 3b respectively, and we can see 13 groups, i.e., two groups are cluttered. Figure 3c and Figure 3d shows the optimal visualization of the Arcviz and CircularViz method. Both of them can display perfectly class preservation. However, the CircularViz shows groups more compactness and more separation than the Arcviz.

Figure 4 shows the optimal CircularViz visualization based on the quality visualization measurement using the KNNC classifier for testing data sets. From Figure 4a to Figure 4f



(a) *RadViz*

(b) *PolyViz*

(c) *ArcViz*

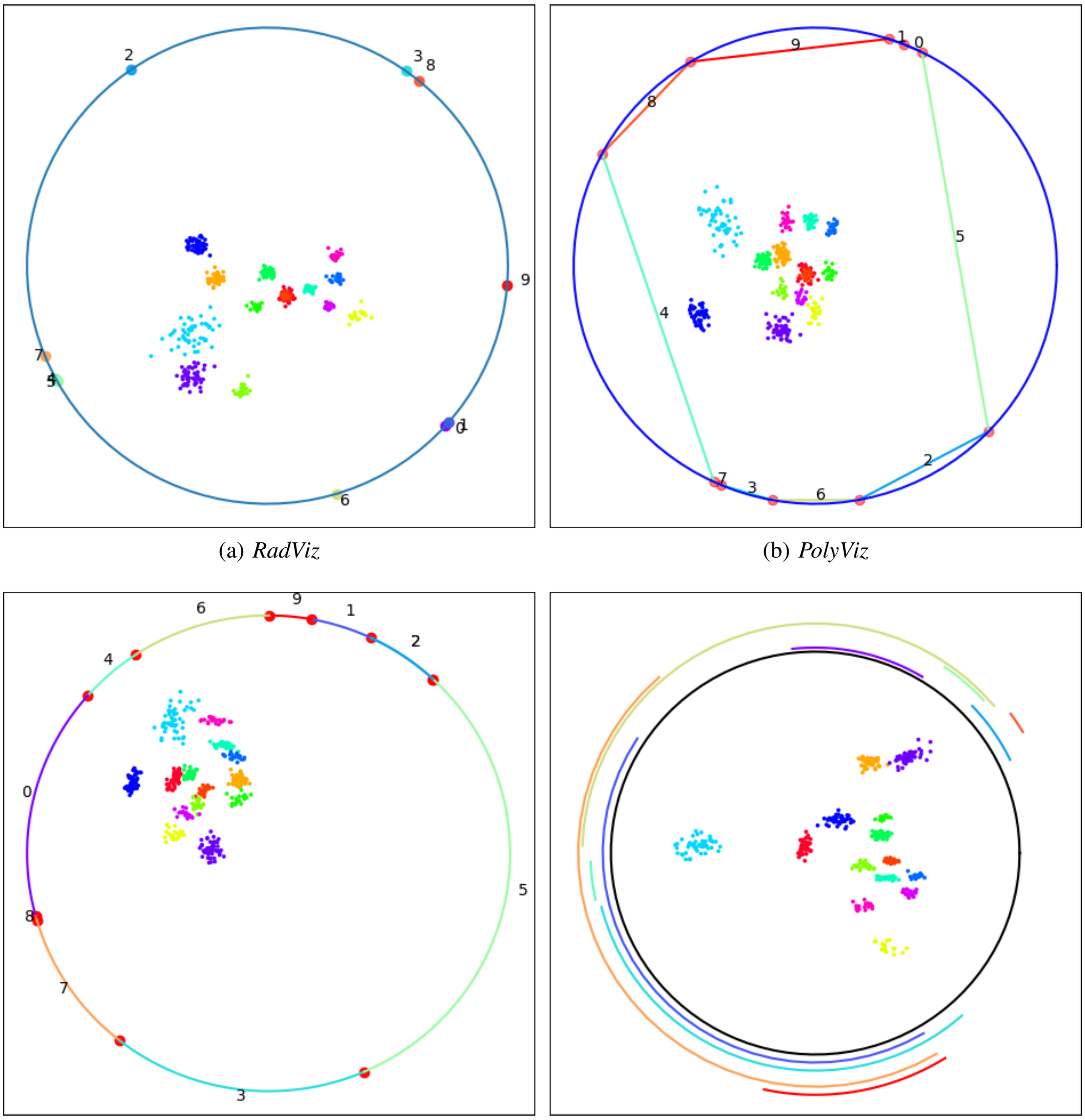
(d) *CircularViz*

Fig. 2. The optimal layout of (a) RadViz, (b) PolyViz, (c) ArcViz, and (d) CircularViz for classes visualization Iris data.

visualizes the Auto-mpg, Ecoli, Iris, Olive, Wine, and Y14c data set. The circular anchors jittered to avoid overlapping on the circle with radius 1.

6. Conclusions and Future Works

We have been presented our approach for supervised data visualization with a point-based method. We propose a modified Radviz method that named as CircularViz method. The CircularViz provides a suitable display of high-dimensional data. We attest the successfulness of the Circularviz method over six experimental data sets and comparison with the Radial visualization, Polygonal visualization, and Arc visualization. For the next development, we investigate some variants of the RadViz methods to enhance the cluster structure of subspace of ultra dimensional data.



(a) *RadViz*

(b) *PolyViz*

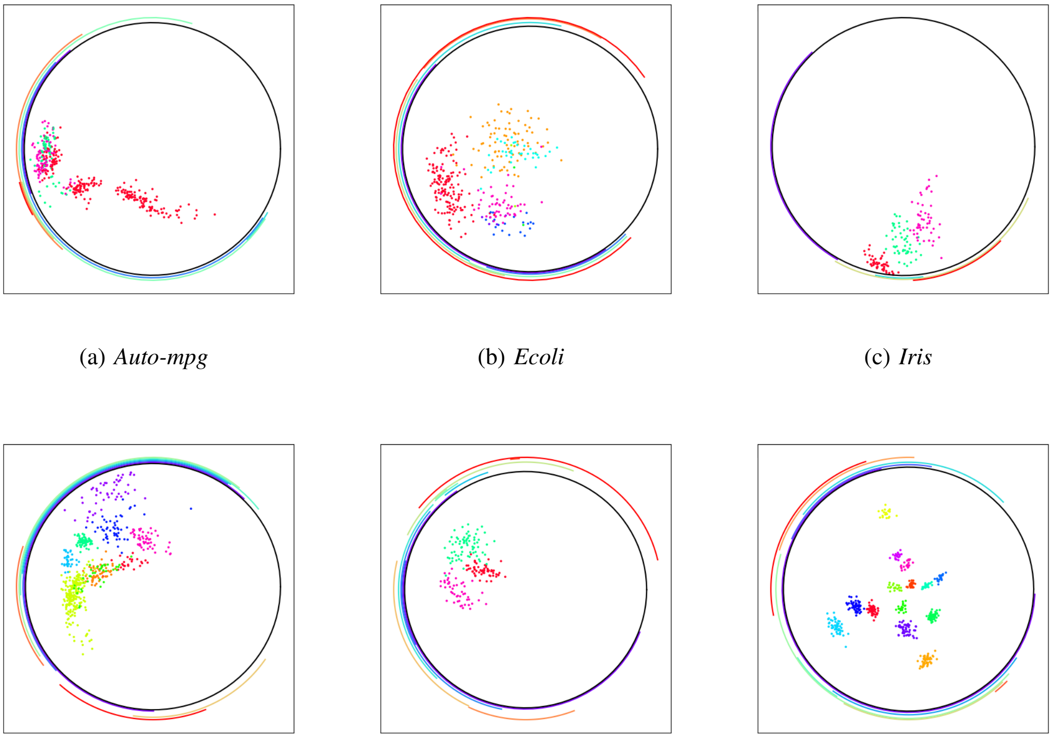
(c) *ArcViz*

(d) *CircularViz*

Fig. 3. The optimal layout of (a) RadViz, (b) PolyViz, (c) ArcViz, and (d) CircularViz for classes visualization of the Y14c data.

Acknowledgment

This research is funded by University of Transport and Communications (UTC) under grant number T2021-CB-010.



(a) *Auto-mpg*

(b) *Ecoli*

(c) *Iris*

(d) *Olive*

(e) *Wine*

(f) *Y14c*

Fig. 4. The optimization CircularViz based on KNNC quality visualization measurements of (a) Auto-mpg, (b) Ecoli, (c) Iris, (d) Olive, (e) Wine, and (f) Y14c data sets.

References

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

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

[3] P. Hoffman, G. Grinstein, K. Marx, I. Grosse, and E. Stanley, “DNA visual and analytic data mining,” in *Proceedings of the 8th conference on Visualization’97*, pp. 437–441, IEEE Computer Society Press, 1997.

[4] P. Hoffman, G. Grinstein, and D. Pinkney, “Dimensional anchors: A graphic primitive for multidimensional multivariate information visualizations,” in *Proceedings of the 1999 workshop on new paradigms in information visualization and manipulation in conjunction with the eighth ACM internation conference on Information and knowledge management*, pp. 9–16, ACM, 1999.

[5] V. L. Tran, “ArcViz: An extended radial visualization for classes separation of high dimensional data,” in *The 10th International Conference on Knowledge and Systems Engineering (KSE 2018)*, pp. 158–162, 2018.

[6] G. Grinstein, M. Trutschl, and U. Cvek, “High-dimensional visualizations,” in *Proceedings of the Visual Data Mining KDD Workshop 2001*, vol. 2, pp. 7–19, ACM, 2001.

[7] J. H. P. Ono, F. Sikansi, D. C. Corrêa, F. V. Paulovich, A. Paiva, and L. G. Nonato, “Concentric Radviz: Visual exploration of multi-task classification,” in *Graphics, Patterns and Images (SIBGRAPI), 2015 28th SIBGRAPI Conference on*, pp. 165–172, IEEE, 2015.

[8] Y. C. Wang, Q. Zhang, F. Lin, C. Goh Keong, and H. S. Seah, “PolarViz: A discriminating visualization and visual analytics tool for high-dimensional data,” *The Visual Computer*, vol. 35, no. 11, pp. 1567–1582, 2019.

[9] I. B. Corrêa and A. C. P. L. F. de Carvalho, “DualRadviz: Preserving context between classification evaluation and data exploration with radviz,” in *Proceedings 5th Brazilian Conference on Intelligent Systems (BRACIS)*, pp. 241–246, IEEE, 2016.

[10] T. Hastie, R. Tibshirani, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction, 2nd Edition (Springer series in statistics).* Springer, 2009.

[11] K. Price, R. M. Storn, and J. A. Lampinen, *Differential evolution: A practical approach to global optimization (Natural Computing Series).* Springer-Verlag New York, Inc., 2005.

[12] M. E. H. Pedersen, “Good parameters for differential evolution,” *Technical Report HL1002, Hvass Laboratories*, 2010.

Manuscript received 11-04-2021; Accepted 16-06-2021.

■

Van Long Tran received the Ph.D. degree in computer science from the Jacobs University, Bremen, Germany, in 2010. He has been involved with academics including teaching and research since 2006. Currently, he is an Associate Professor at University of Transport and Communications, Hanoi, Vietnam. He is doing research in the field of data visualization, specialized with information visualization. E-mail: vtran@utc.edu.vn

CIRCULARVIZ: TRỰC QUAN HÓA NÂNG CAO CHO DỮ LIỆU NHIỀU CHIỀU

Trần Văn Long

Tóm tắt

Trực quan hóa cho dữ liệu nhiều chiều là lĩnh vực nghiên cứu quan trọng trong khám phá dữ liệu và được ứng dụng trong nhiều lĩnh vực khoa học kỹ thuật. Phương pháp Radviz là một trong các phương pháp phổ biến để trực quan hóa dữ liệu nhiều chiều. Phương pháp Radviz cho phép trực quan hóa về cấu trúc của dữ liệu nhiều chiều. Các phương pháp truyền thống của Radviz là tìm tối ưu thứ tự sắp xếp các chiều trên đường tròn đơn vị để biểu diễn cấu trúc dữ liệu nhiều chiều. Trong bài báo này, chúng tôi giới thiệu về một cải tiến của phương pháp chiếu Radviz để thể hiện cấu trúc nhóm của dữ liệu trong không gian nhiều chiều. Phương pháp tiếp cận trong Radviz cải tiến là thay vì biểu diễn số chiều bởi một điểm bằng một cung trên đường tròn đơn vị và mỗi cung tròn đó được biểu diễn dữ liệu của thuộc tính. Phương pháp cải tiến nhằm mục tiêu mở rộng không gian biểu diễn và bảo toàn cấu trúc nhóm của dữ liệu. Phương pháp tiếp cận mới được chứng minh với hai loại độ đo chất lượng của phương pháp biểu diễn với một số dữ liệu nhiều chiều.