Another Look at Radial Visualization for Class-preserving Multivariate Data Visualization

Van Long Tran
University of Transport and Communications, Hanoi, Vietnam
E-mail: vtran@utc.edu.vn

Keywords: data visualization, radial visualization, quality visualization

Received: March 24, 2017

Multivariate data visualization is an interesting research field with many applications in various fields of sciences. Radial visualization is one of the most common information visualization concept for visualizing multivariate data. However, radial visualization may display different information about structures of multivariate data. For example, all points which are multiplicatives of given point may map to the same point in the visual space. An optimal layout of radial visualization is usually found by defining a suitable the order of data dimensions on the unit circle. In this paper, we propose a novel method that improves the radial visualization layout for cluster preservation of multivariate data. The traditional radial visualizations have viewpoint at the origin coordinate. The idea of our proposed method is finding the most suitable viewpoint among the corners of a hypercube to look into the cluster structures of data sets. Our method provides an improvement in visualizing class structures of multivariate data sets on the radial visualization. We present our method with three kinds of quality measurements and prove the effectiveness of our method for several data sets.

Povzetek: Predstavljena je vizualizacija multivariantnih podatkov.

1 Introduction

Many scientific and business applications produce large data sets with increasing complexity and dimensionality. While information is growing in an exponential way, data are ubiquitous in our world. Data should contain some kind of valuable information that can possibly be explored using human knowledge. However, extracting meaningful information in large scale data is a difficult task.

Information visualization techniques have been proven to be of high value in gaining insight into these large data sets. The aim of information visualization is to use the computer-based interactive visual representations of abstract and non-physically based data to amplify human cognition. It aims at helping users to detect effectively and explore the expected, as well as discovering the unexpected, to gain insight into the data [6].

A major challenge for information visualization is how to present multidimensional data to analysts, because complex visual structures occur. Data visualization methods often employ a map from multidimensional data into lower-dimensional visual space. The reason is that visual space representation is composed of two or three spatial coordinates and a limited number of visual factors such as color, texture, etc. However, when the dimensionality of the data is high, usually from tens to hundreds, the mapping from multidimensional data space into visual space imposes information loss. This leads to one of the big question in information visualization [6]: How to project from a multidimensional data space into a low-dimensional space and best preserve the characteristics of the data.

The order of data dimensions is a crucial problem for the effectiveness of many multidimensional data visualization techniques [3] such as parallel coordinates [13], star coordinates [14], Radial visualization (Radviz) [10], scatterplot matrix [2], circle segments [4], and pixel recursive pattern [15]. The data dimensions have to be positioned in some one- or two- dimensional arrangement on the screen. The chosen arrangement of data dimensions can have a major impact on the expressiveness of the visualization because the relationships among adjacent dimensions are easier to detect than relationships among dimensions positioned far from each other. Dimension ordering aims to improve the effectiveness of the visualization by giving reasonable orders to the dimensions so that users can easily detect relationships or pay more attention to more important dimensions.

The Radviz technique is one of the most common visualization techniques used in medical analysis [10, 11, 16]. Finding the optimal order of data dimensions in Radviz is known to be NP-complete [3]. Although there have been a number of proposed methods for solving the dimension ordering problem in Radviz [16, 8], most of them are exhaustive or greedy local searches in the space of all permutations of data dimensions. These methods are usually only tested on some data sets with small number of dimensions.

One of the disadvantages of Radviz is that all multidimensional points which differ by a multiplicative constant, i.e., all points cp with a fixed point p and various non-zero scalars c, number that map to the same position in the visual space. Thus, all these points separate in the original space but they cannot be differentiated in the visual space. This property is invariant for all permutations. Radviz can be explained as a combination of a perspective projection and a linear mapping with the viewpoint at the origin and the view plane being a simplex. In this paper, we propose another variant of Radviz that supports users visualizing the data inside a hypercube from an arbitrary viewpoint at the corners of the hypercube. Finding a suitable viewpoint of the hypercube in an n-dimensional space has 2^n possible cases. In general, finding a good viewpoint is less complicated than finding a good data dimensions permutation of Radviz.

The remaining part of this paper is organized as follows. In Section 2, we present related work with Radviz and data dimensions reordering in multivariate data visualization techniques. The inversion axes in Radviz are presented in Section 3. In Section 4, we describe some methods for measurement quality of class visualizations for multivariate data in the visual space. In Section 5, we show the effectiveness of our methods with five well known multivariate data sets in the case of classified data. In Section 6, we make a comparison for five data sets with permutations in Radviz with other algorithms. In Section 7, we present our conclusion and future work.

2 Related work

Principal Component Analysis (PCA) is one of the most common methods for the analysis of multivariate data [12]. PCA is applied to visualizing multivariate data that is a linear projection onto two or three eigenvectors. The general linear mapping can be defined as $P(x) = Vx$ where V is a matrix. PCA projects a multidimensional point x into a space spanned by the two or three eigenvectors that corresponding to the two or three largest eigenvalues of the covariance matrix of the given data sets.

Star coordinates were introduced by Kandogan [14]. Star coordinates use a linear mapping with the ith column of matrix transformation $V_i = (\cos\frac{2\pi i}{n}, \sin\frac{2\pi i}{n})^T$. Vectors $\{V_i, i = 1, 2, \ldots, n\}$ are represented evenly on the unit circle in the two-dimensional visual space. The author also introduced several techniques for interactions on star coordinates, for example moving axes V_i in the visual space. In [5], 3D star coordinates are introduced with $V_i = (\cos\frac{2\pi i}{n}, \sin\frac{2\pi i}{n}, 1)^T$ that extends the 2D star coordinates by adding the third coordinates as summation of all coordinates. Further properties can be found in [20, 17]. Long and Linsen [22] propose optimal 3D star coordinates for visualizing hierarchical clusters in multidimensional data.

Radviz was proposed by Hoffman et al. [10]. Radviz can be explained as a perspective projection of the 3D star coordinates with a view point at the origin and viewing plane $z = 1$. A normalized Radviz and properties of Radviz are presented in [7]. The important problem with Radviz is the ordering of the dimensional anchors for a good viewing of the multivariate data. In [19], the t-statistic method for reordering dimensional anchors on the unit circle is introduced. The t-statistic is applied for labelled data. Di Caro et al. [8] proposed two methods for dimension arrangement in Radviz based on an optimization problem for pair of similarity matrix between data dimensions and neighbourhood matrix between data dimensions on a unit circle [8]. Albuquerque et al. [1] used the Cluster Density Measure (CDM) for finding a good layout of Radviz. The authors propose a greedy incremental algorithm to successively add data dimensions to the Radviz layout to determine a suitable order.

3 Radial visualization method

3.1 Radviz

Radviz was first introduced by Hoffman et al. in [10, 11], and it could be regarded as an effective non-linear dimensionality reduction method. Radviz directly maps multidimensional data points into a visual space based on an equibalance of spring systems. In Radviz, dimensional anchors are attached to springs. The stiffness of each spring equals the value of the dimension corresponding to its dimensional anchor. The other end of each spring is attached to a point in the visual space. The location of this point ensures the equibalance of the spring systems.

Let $x = (x_1, x_2, \ldots, x_n)$ be a data point in a hypercube $[0, 1]^n$. The dimensional anchors $S_i, i = 1, 2, \ldots, n$ can be easily calculated by the formula:

$$S_i = (\cos\frac{2\pi(i-1)}{n}, \sin\frac{2\pi(i-1)}{n}), i = 1, 2, \ldots, n.$$

For the spring systems to be equibalanced, we must have $\sum\limits_{i=1}^{n} x_i(p - x_i) = 0$, and we have the location of p as follows:

$$p = \frac{\sum_{i=1}^{n} x_i S_i}{\sum_{i=1}^{n} x_i}. \quad (1)$$

Thus, the multidimensional point x is represented by the point p. Figure 1 shows how a sample x of an eight-dimensional space is represented by a point p in a 2-dimensional plot.

The important properties of the Radviz method are described in [7]:

- If a multidimensional point with all x coordinates have the same value, the data point lies exactly in the origin of the graph. Points with approximately equal dimensional values (after normalization) lie close to the center. Points with similar dimensional values, whose dimensions anchors are opposite each other on the circle lie near the center.

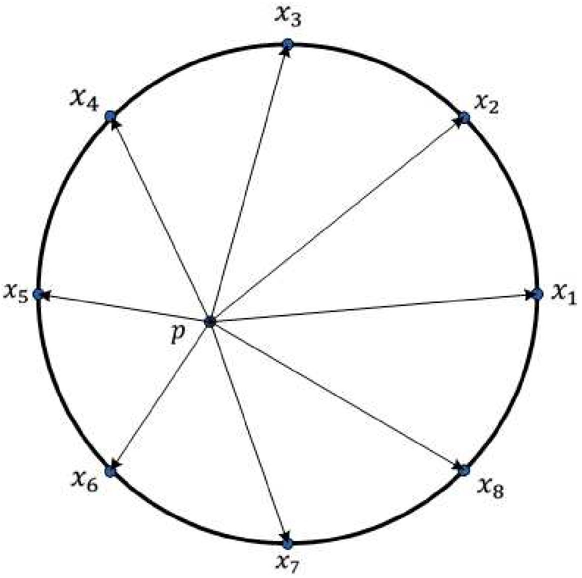

Figure 1: Radviz visualizes a point in 8 dimensions. The dimensions are represented by points, placed equally spaced on the unit circle. An observation x is displayed at position p corresponding to its attributes $x_1, x_2, \ldots, x_8$.

- If the point is a unit vector point, it lies exactly at the fixed point on the edge of the circle where the spring for that dimension is fixed. Points which have one or two coordinate values significantly greater than the others lie closer to the dimensional anchors (fixed points) of those dimensions.
- The position of a point depends on the layout of the particular dimensional anchors around the circle.
- Many points can be mapped to the same position. This mapping represents a non-linear transformation of the data that preserves certain symmetries.
- The Radviz method maps each data record to a point in a multidimensional data set that is within the convex hull of the dimensional anchors.

We can consider the Radviz nonlinear mapping as a combination of a perspective projection with the viewer at $o = (0, 0, \ldots, 0)$ on a simplex $\sum_{i=1}^{n} x_i = 1$, $V(x) = (\sum_{i=1}^{n} x_i)^{-1} x$ and a linear mapping as in the Star coordinates [14] $L_S(x) = \sum_{i=1}^{n} x_i S_i$. The Radviz mapping can be rewritten as follows:

$$R(x) = L_S(V(x)) = (\sum_{i=1}^{n} x_i)^{-1} \sum_{i=1}^{n} x_i S_i.$$

3.2 Inversion Radviz

We propose a method that supports users in viewing the hypercube at arbitrary corner of the unit hypercube. We assume that the view is a point $p = (p_1, p_2, \ldots, p_n) \in \{0,1\}^n$. The simplex at the point p is a hyperplane (π_p) that goes through n points $(p_1, \ldots, 1 - p_i, \ldots, p_n), i = 1, 2, \ldots, n$. The equation of the simplex is determined as follows:

$$\sum_{i=1}^{n} (1 - 2p_i) x_i = 1 - \sum_{i=1}^{n} p_i.$$

We can rewrite the above equation of the hyperplane as

$$(\pi_p) : \sum_{p_i=0} x_i + \sum_{p_i=1} (1 - x_i) = 1.$$

We find the position of the multidimensional point $x = (x_1, x_2, \ldots, x_n) \in [0,1]^n$ in the visual space. The coordinates of the point x with respect to the origin p and the basic vectors

$$\Big((1 - 2p_1)e_1, (1 - 2p_2)e_2, \ldots, (1 - 2p_n)e_n\Big),$$

is denoted by

$$x_p = (\frac{x_1 - p_1}{1 - 2p_1}, \frac{x_2 - p_2}{1 - 2p_2}, \ldots, \frac{x_2 - p_2}{1 - 2p_n}),$$

or

$$x_p = \Big(p_1 + (1 - 2p_1)x_1, \ldots, p_n + (1 - 2p_n)x_n\Big),$$

where $(e_1, e_2, \ldots, e_n)$ are the standard basic vectors of R^n. Obviously, the coordinates of the point x are the coordinates of the vector $x - p$ with respect to the vector basic systems above.

The perspective projection V maps a point x_p onto the hyperplane (π_p) at the point $V_p(x)$ where

$$V_p(x) = \frac{\Big(p_1 + (1 - 2p_1)x_1, \ldots, p_n + (1 - 2p_n)x_n\Big)}{\sum_{p_i=0} x_i + \sum_{p_i=1} (1 - x_i)}.$$

Figure 2 displays the viewpoint p, the view plane (π_p), and the location $V_p(x)$ of the multidimensional point x on the hyperplane (π_p).

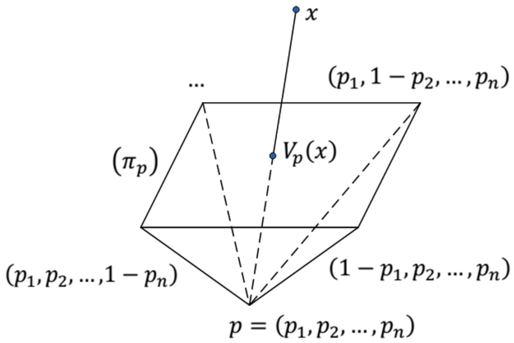
x
...
$(p_1, 1 - p_2, \ldots, p_n)$
(π_p)
$V_p(x)$
$(p_1, p_2, \ldots, 1 - p_n)$
$(1 - p_1, p_2, \ldots, p_n)$
$p = (p_1, p_2, \ldots, p_n)$

Figure 2: The perspective projection at corner p.

The Radviz projection at the point p is defined as

$$P(x) = \frac{\sum_{i=1}^{n} \Big(p_i + (1 - 2p_i)x_i\Big) S_i}{\sum_{p_i=0} x_i + \sum_{p_i=1} (1 - x_i)},$$

or

$$P(x) = \frac{\sum\limits_{p_i=0} x_i S_i + \sum\limits_{p_i=1} (1 - x_i) S_i}{\sum\limits_{p_i=0} x_i + \sum\limits_{p_i=1} (1 - x_i)}.$$

The ith coordinate of the point x corresponding to $p_i = 1$ is changed to $1 - x_i$. We propose an inversion Radviz (iRadviz for short) to project the multidimensional point x onto the visual space as follows:

$$R_{p,S}(x) = \frac{\sum\limits_{p_i=0} x_i S_i + \sum\limits_{p_i=1} (1 - x_i) S_i}{\sum\limits_{p_i=0} x_i + \sum\limits_{p_i=1} (1 - x_i)} \qquad (2)$$

Figure 3 shows the Radviz and iRadviz to visualize a synthetic data set in three dimensional space that called as 3D data set. The 3D data set contains 700 points which split into seven clusters. Each cluster has 100 points at the seven vertices of the cube except vertex $(1, 1, 1)$. Figure 3 (left) shows the traditional Radviz visualizing the 3D data set. One cluster at the origin $(0, 0, 0)$ is spread on the simplex. Radviz visualizes three dimensional space data set that is not affected by permutation. Figure 3 (right) shows the 3D data set with iRadviz using viewpoint $(1, 1, 1)$ where the seven clusters are perfectly separated.

For interaction, users can select a dimensional anchor p_i in Radviz and change this vertex into $1 - p_i$. For finding the optimal viewpoint of the iRadviz of the given data set, we need a quality measurement to define a suitable view of a multidimensional data set.

4 Quality measurement

Suppose data set $X = \{\mathbf{x}_i : 1 \leq i \leq n\}$ is classified into K classes and each class is labeled by $C = \{1, 2, \ldots, K\}$. We denote n_k a the number of data points in the kth class. In this section, we present briefly three methods to measure quality in iRadviz for visualizing supervised data. Without loss of generality, we also denote the data set that is projected in the visual space by $X = \{x_i : 1 \leq i \leq n\} \subset R^2$.

4.1 Class distance consistency

For each class, we denote $\mathbf{c}_k$ as the centroid of the k^{th} class. A data point $\mathbf{x}$ belongs to a particular class if the distance from the data point $\mathbf{x}$ to the centroid of this class is smallest. Hence, we denote

$$class(\mathbf{x}) = \arg \min_{1 \leq k \leq K} ||\mathbf{x} - \mathbf{c}_k||.$$

A data point $\mathbf{x}$ is correctly represented if its label is the same as its class, otherwise the data point $\mathbf{x}$ a miss.

The Class Distance Consistency (CDC) [21] of data set $X = \{x_i : 1 \leq i \leq n\}$ is defined as the number of correctly represented data points, i.e.,

$$Q(\text{CDC}, X) = \frac{|\mathbf{x}_i : label(\mathbf{x}_i) = class(\mathbf{x}_i)|}{n}. \qquad (3)$$

The CDC quality measurement for class visualization is applicable for a spherical shape of clusters.

4.2 Cluster density measurement

The quality Cluster Density Measurement (CDM) [1] is defined as follows:

$$Q(\text{CDM}, X) = \sum_{i,j=1}^{K} \frac{d_{ij}^2}{r_i r_j}, \qquad (4)$$

where $d_{ij} = ||c_i - c_j||$ is the Euclidean distance between two cluster centroids, and r_i is an average radius of the ith cluster, i.e.,

$$r_i = \frac{\sum\limits_{label(x)=i} ||x - c_i||}{n_i}.$$

The high value quality presents well defined cluster separations with small intra-cluster distances and large inter-cluster distances. Hence, the higher the quality measure is, the better is the visualization of the supervised data set.

4.3 Conditional entropy

The Havrda-Charvat's structural α-entropy [9] is defined as

$$H_\alpha(X) = \frac{2^{\alpha-1}}{2^{\alpha-1} - 1}\Big(1 - \sum_{i=1}^{n} p^\alpha(x_i)\Big), \alpha > 0, \alpha \neq 1.$$

A conditional Havrda-Charvat's structural α-entropy [18] for class visualization quality is defined as follows:

$$\begin{aligned} H_\alpha(C|X) &= \int p(x) H_\alpha(C|X = x) dx \\ &= \frac{2^{\alpha-1}}{2^{\alpha-1} - 1}\Big(1 - \sum_{j=1}^{K} \int p^\alpha(j|x) p(x) dx\Big). \end{aligned}$$

We can estimate the conditional entropy $H_\alpha(C|X)$ as follows:

$$H_\alpha(C|X) = \frac{2^{\alpha-1}}{2^{\alpha-1} - 1}\Big(1 - \frac{1}{n}\sum_{j=1}^{K}\sum_{i=1}^{n} p^\alpha(j|x_i)\Big).$$

Assume each data point x_i is classified into only one class, i.e., $p(j|x_i) = 1$ for the jth class and $p(j|x_i) = 0$ for any other class. The conditional entropy achieves minimal value.

When $\alpha = 2$, we have the quadratic entropy:

$$H_2(C|X) = 2\Big(1 - \frac{1}{n}\sum_{j=1}^{K}\sum_{i=1}^{n} p^2(j|x_i)\Big).$$

By Bayes' theorem, we have

$$p(j|x) = \frac{p(j) p(x|j)}{p(x)}.$$

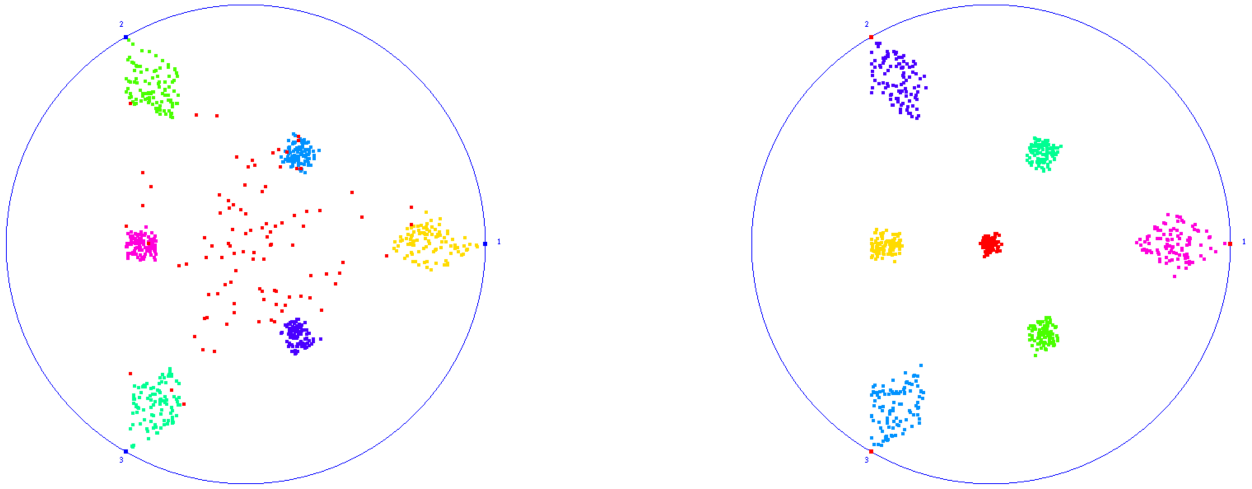

Figure 3: The synthetic 3D data visualization. *(Left)* Traditional Radviz. *(Right)* iRadviz with viewpoint $(1, 1, 1)$.

The prior probability is estimated by

$$p(j) = \frac{n_j}{n}.$$

The density $p(x|j)$ and $p(x)$ are estimated by nonparametric techniques as the Parzen window method. Consider a small region $R(x)$ that contains x and has area V. Assume the region $R(x)$ contains $k_j(x)$ points of the jth class and $k(x)$ points of the data set. We estimate the density by

$$p(x|j) = \frac{k_j(x)}{n_j V},$$

and $p(x) = \dfrac{k(x)}{nV}$. Therefore, the conditional probability $p(j|x)$ can be estimated by

$$p(j|x) = \frac{\dfrac{n_j}{n}\dfrac{k_j(x)}{n_j V}}{\dfrac{k(x)}{nV}} = \frac{k_j(x)}{k(x)}.$$

The quality entropy is defined as following

$$Q(ENT, X) = 1 - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K}\left(\frac{k_j(x_i)}{k(x_i)}\right)^2 \qquad (5)$$

The lower the quality entropy is, the better is the clustering visualization. For calculating the entropy quality, we divide the square region that contains all data set into $N \times N$ grid cells. The grid size N in two-dimensional space is estimated by the k-nearest neighbor. For each cell c, we have 9 neighbor cells, and on average in 9 cells we have $\dfrac{9n}{N^2}$ points. The grid size N is calculated by $\dfrac{9n}{N^2} = \sqrt{n}$ or

$$N = 1 + \left[3\sqrt[4]{n}\right].$$

For each cell c, we store the class point counts $c = (c_1, c_2, \ldots, c_K)$, where c_j is the number of point of the jth class falling into the cell c. For each point x that falls in the cell c, region $R(x)$ contains all cells that are neighbors with the cell c. We have $k_j(x) = \sum\limits_{c \in R(x)} c_j$ and $k(x) = \sum\limits_{c \in R(x)} k_j(x)$. The complexity for computing the entropy quality is $O(Kn)$, i.e., it has linear time complexity.

5 Experimental results

We tested our approach with five data sets. For each data set, we find the viewpoint for the iRadviz based on the three quality measurements presented in the Section 4.

The first well known data set is called the Iris data set[1]. The Iris data set contains 150 data points, four attributes: X_1 (sepal length), X_2 (sepal width), X_3 (petal length), X_4 (petal width) and three classes: Setosa (50 data points), Versicolour (50 data points), and Virginica (50 data points).

Figure 4 shows the iRadviz approach for visualizing the Iris data set. Classes are encoded by different colors. One class (red) is separated perfectly with two other classes. In Figure 4 (left) with inversion of the axes X_2, X_3, X_4 and Figure 4 (right) with inversion of the axes X_1, X_2, X_3, X_4. These figures show three classes better separated than in Figure 4 (middle) without inversion the axes.

The second data set is named the Wine data set[2]. The Wine data set includes 178 data points with 13 attributes: X_1(Alcohol), X_2 (Malic acid), X_3 (Ash), X_4 (Alcalinity of ash), X_5 (Magnesium), X_6 (Total phenols), X_7 (Flavanoids), X_8 (Nonflavanoid phenols), X_8 (Proanthocyanins), X_{10} (Color intensity), X_{11} (Hue), X_{12} (OD280 /

[1] http://archive.ics.uci.edu/ml/datasets/Iris

[2] http://archive.ics.uci.edu/ml/datasets/Wine

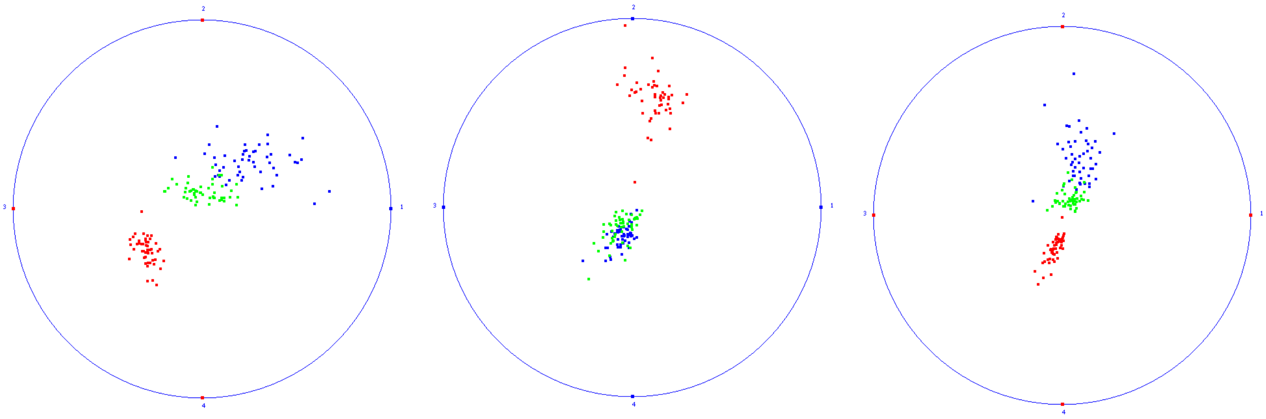

Figure 4: The Iris data. *(Left)* The best iRadviz visualization based on CDC quality. *(Middle)* The best iRadviz visualization based on CDM quality. *(Right)* The best iRadviz visualization based on Entropy quality.

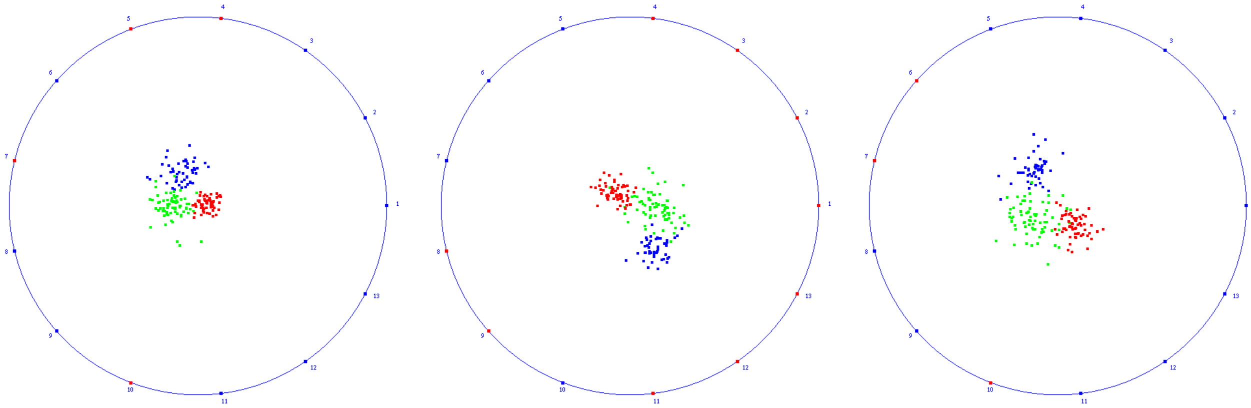

Figure 5: The Wine data. *(Left)* The best CDC quality of iRadviz visualization. *(Middle)* The best quality CDM of iRadviz visualization. *(Right)* The best quality Entropy of iRadviz visualization.

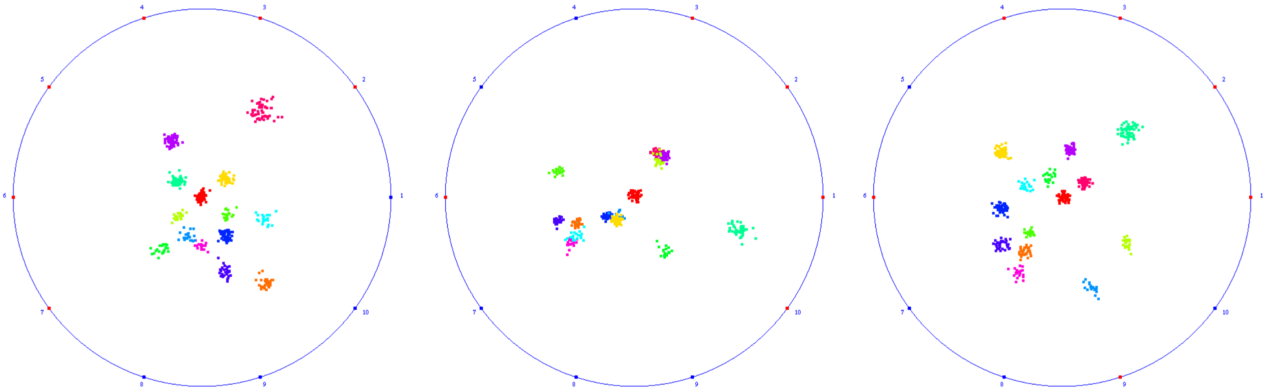

Figure 6: The Y14c data. *(Left)* The best quality CDC on iRadviz. *(Middle)* The best quality CDM on iRadviz. *(Right)* The best quality Entropy on iRadviz.

OD315 of diluted wines), and X_{13} (Proline). The Wine data set is classified into three classes: class 1 (59 data points), class 2 (71 data points), and class 3 (48 data points). Figure 5 shows the Wine data set with a different viewpoint using iRadviz. The different colors represent different classes of the Wine data set. Figure 5 (left) shows the best iRadviz visualization for the Wine data set with highest CDC quality where inversion was applied to axes X_4, X_5, X_7, X_{10}. Figure 5 (middle) shows the best iRadviz visualization for the wine data set with highest CDM quality where inversion has been applied to axes $X_1, X_2, X_3, X_4, X_8, X_9, X_{11}, X_{12}, X_{13}$. Figure 5 (right) shows the best iRadviz visualization for the wine data set with highest Entropy quality where inversion has been applied to axes X_6, X_7, X_{10}.

The third data set is a synthetic data set, that contains 480 data points with ten attributes and partitions into 14 clusters. Figure 6 shows three views of the Y14c data with several different viewpoints in iRadviz. In this figure, the inversion axes are highlighted by red color. Figure 6 (left) shows the best iRadviz class visualization of this data on the CDC quality with inversion axes $2, 3, 4, 5, 6, 7$. Clusters shown in this figure are well separated. Figure 6 (middle) shows the best iRadviz based on highest CDM quality with inversion axes $1, 2, 3, 6, 10$. Several clusters are overlapping in this visualization. Figure 6 (right) shows the best iRadviz based on highest Entropy quality with inversion axes $1, 2, 3, 4, 6, 9$. This figure shows that clusters are perfectly separated. The Y14c data set contains two clusters that have an different a scale. These clusters are fully overlapped on the Radviz with any permutation of dimensional anchors.

The fourth data set is named Italian Olive Oils data (Olive for short)[3]. The Olive data set consists of eight attributes about eight fatty acids (X_1 palmitic, X_2 palmitoleic, X_3 stearic, X_4 oleic, X_5 linoleic, X_6 linolenic, X_7 arachidic, X_8 eicosenoic) and 572 data samples. The Olive data set is classified into nine clusters. Each cluster corresponds to one of nine areas in Italy. Figure 7 shows the iRadviz class visualization of the Olive data set that shows the best quality based on CDC (left), CDM (middle), and Entropy (right). Figure 7 (left and right) classes are more separated than the classes in Figure 7 (middle).

The last data set is called Ecoli[4]. The Ecoli data contains 336 data samples and each data sample consists seven attributes. The Ecoli data set is partitioned into eight clusters with $143, 77, 52, 35, 20, 5, 2, 2$ data samples respectively. The three last clusters contain very small data amounts of samples. Figure 8 shows the class visualization using iRadviz with the best quality based on CDC (left), CDM (middle), and Entropy (right).

6 Comparison and discussion

In this section, we present some quality measurements of our proposed method versus permutation and our method versus other algorithms.

6.1 Inversion dimension versus permutation

For the three first data set (Iris, Ecoli, and Olive) data sets, we find the global best permutation for each quality measurements by searching over all permutations. The two last data sets (Y14c and Wine), we find the local best permutation. We call two instances permutations of data dimension if they are different by one consecutive position. The local best permutation achieves the best quality over all neighbor permutations.

Class Distance Consistency: Table 1 shows that the quality of our approach is better than the CDC quality in [21] for the Iris, Ecoli, Y14c, and Wine data sets and is slightly lower than the CDC quality for the Olive data set.

Cluster Density Measurement: Table 2 shows that the CDM quality of our approach is better than the CDM quality in [2] for the two last data sets, lower for the Ecoli and Olive data sets, and the same for the Iris data set.

Entropy Measurement: Table 3 shows that the Entropy quality of our approach is better than the Entropy quality in [18] for the Iris, Ecoli, and Y14c data sets, and is slightly lower for the Olive and Wine data sets.

6.2 Inversion axes with other permutation algorithms

In this section, we present the quality measurement comparison of our method versus the t-statistic method and the CDM method about the permutation on the Radviz [1]. The best permutation in Radviz for the Wine data by t-statistic method is $\{1, 2, 4, 8, 10, 11, 13, 12, 9, 7, 6, 5, 3\}$, and the CDM method delivers $\{8, 3, 4, 2, 10, 13, 1, 5, 6, 7, 9, 12, 11\}$. The best permutation in Radviz for the Olive data by t-statistic method is $\{1, 2, 5, 4, 8, 7, 3, 6\}$, and the CDM method delivers $\{1, 3, 4, 7, 6, 2, 8\}$.

Table 4 shows the quality measurements CDC, CDM, and Entropy (ENT) for the Olive and Wine data sets. The overall quality measurements of our approach are better than those of the t-statistic and CDM methods except for the Entropy quality measure applied to the Wine data set.

Figure 9 (left) shows Radviz visualizing the Wine data set with the best permutation by the t-statistic method and Figure 9 (right) shows the Radviz visualizing the Wine data set with the best permutation by the CDM method. In comparison, Figure 5 shows the Wine data set over the inversion axes. The Figure 9 (left) shows lowest quality for class separation for the Wine data set, while Figure 5 (left) shows the highest quality for class separation.

Figure 10 shows the Olive data set with the two best permutations using the t-statistic method (left) and the CDM method (right). Comparison with the inversion axes layout is provided in Figure 7. Figure 7 (left) and Figure 10

[3] http://cran.r-project.org/

[4] https://archive.ics.uci.edu/ml/datasets/Ecoli

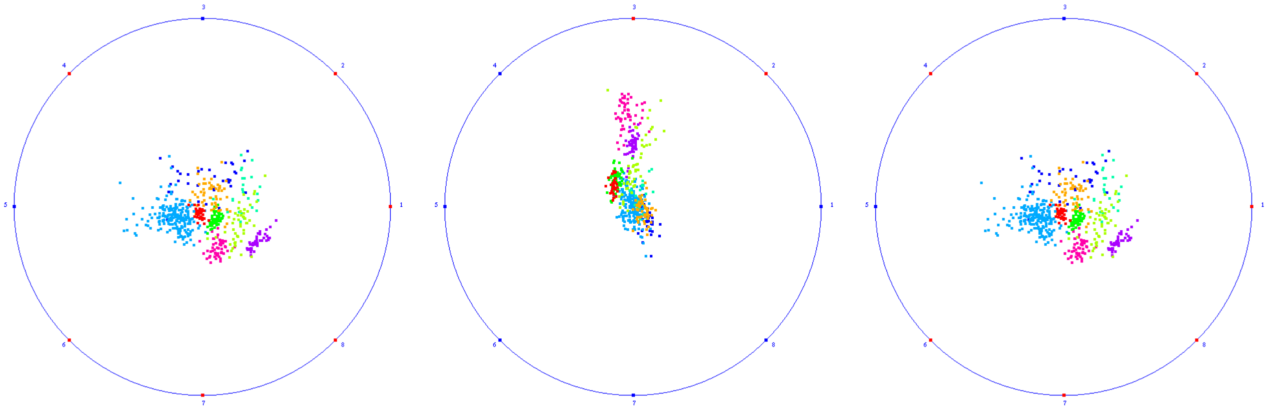

Figure 7: The Olives Oil data. *(Left)* The best quality CDC on iRadviz. *(Middle)* The best quality CDM on Radviz. *(Right)* The best quality Entropy on iRadviz.

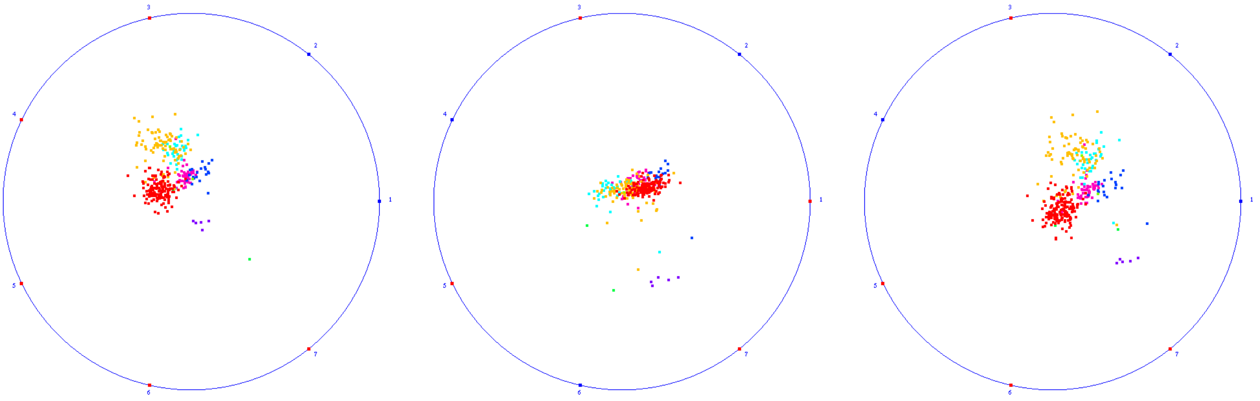

Figure 8: The Ecoli data set. *(Left)* The best quality CDC on iRadviz. *(Middle)* The best quality CDM on iRadviz. *(Right)* The best quality Entropy on iRadviz.

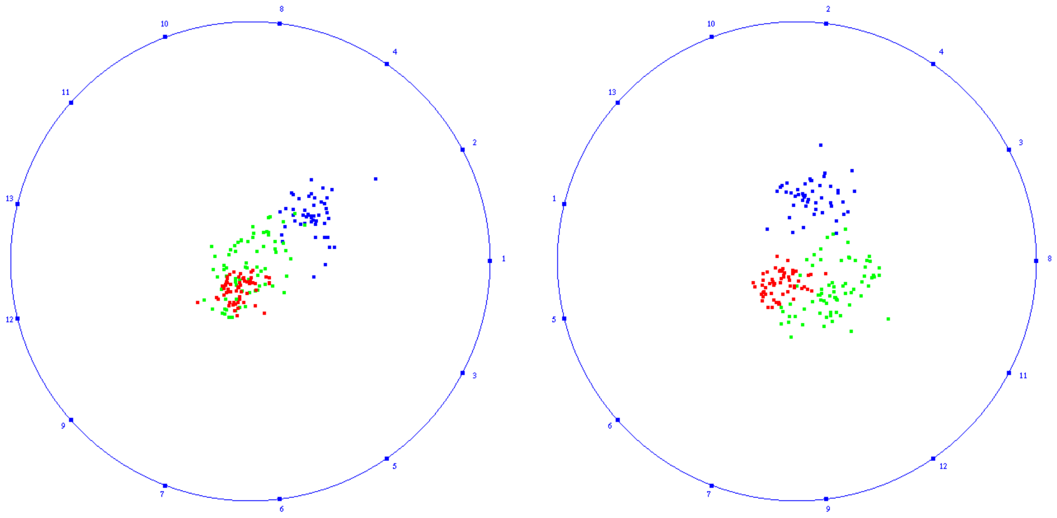

Figure 9: The Wine data. *(Left)* The best permutation by t-statistic method. *(Right)* The best permutation by CDM method.

CDC	Iris	Ecoli	Olive	Y14c	Wine
Permutation	84.67%	67.56%	82.34%	93.96%	94.94%
iRadviz	94.00%	78.57%	80.24%	100%	96.63%

Table 1: The best CDC function over permutation and inversion axes.

Quality CDM	Iris	Ecoli	Olive Oil	Y14c	Wine
Permutation	44.242	42.457	27.825	358.37	13.914
iRadviz	44.242	32.325	23.078	459.824	16.634

Table 2: The Best CDM function over permutation and inversion axes.

Entropy	Iris	Ecoli	Olive Oil	Y14c	Wine
Permutation	0.1316	0.2057	0.1198	0.0648	0.0084
iRadviz	0.0028	0.1645	0.1281	0.000	0.0261

Table 3: The Best Entropy function over permutation and inversion axes.

Data	Olive		Wine	
Method	CDC	ENT	CDC	ENT
t-statistic	55.95%	0.4090	75.28%	0.1643
CDM	76.57%	0.1826	88.87%	0.0176
Our method	80.02%	0.1281	96.63%	0.0261

Table 4: The quality measurement for the Olive and Wine data sets.

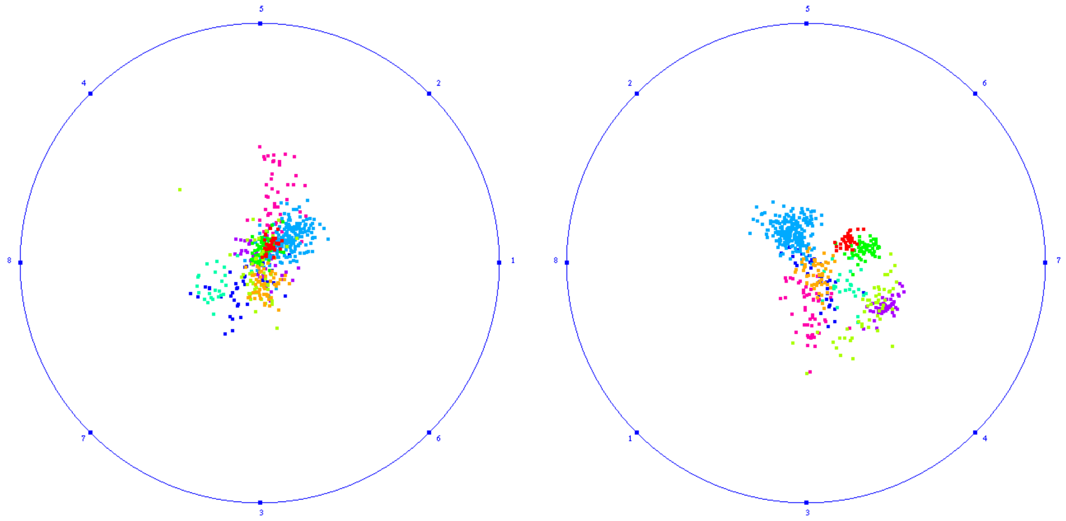

Figure 10: The Olives Oil data. *(Left)* The best permutation with CDC quality. *(Right)* The best permutation with Entropy quality.

(right) have the lowest quality for class separation in the visual space while Figure 7 (left and right) exhibits higher quality for class separation for both permutations.

7 Conclusion

We have presented a new method for visualizing multidimensional data based on Radial visualization. Our proposed method supports users choosing a suitable view for data sets in hypercube. We proved the effectiveness of our method versus permutation dimensional anchors on the Radviz for some supervised data both synthetic and real.

For future work, we want to improve our method to enhance class structures in subspaces with supervised data sets. Moreover, we want to develop other quality measurements for supervised data sets.

Acknowledgement

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.01-2012.04.

References

[1] G. Albuquerque, M. Eisemann, D. J. Lehmann, H. Theisel, and M. Magnor. Improving the visual analysis of high-dimensional datasets using quality measures. In *IEEE Symposium on Visual Analytics Science and Technology (VAST), 2010*, pages 19–26, 2010.

[2] G. Albuquerque, M. Eisemann, D. J. Lehmann, H. Theisel, and M. A. Magnor. Quality-based visualization matrices. In *Proceedings of the Vision, Modeling and Visualization Workshop 2009 (VMV), Braunschweig, Germany*, pages 341–350, 2009.

[3] M. Ankerst, S. Berchtold, and D. A. Keim. Similarity clustering of dimensions for an enhanced visualization of multidimensional data. In *Proceedings IEEE Symposium on Information Visualization (InfoVis '98), 1998*, pages 52–60, 1998.

[4] M. Ankerst, D. A. Keim, and H.-P. Kriegel. Circle segments: A technique for visually exploring large multidimensional data sets. *Proceedings of the 1996 IEEE Symposium on Information Visualization, Hot Topic Session, San Francisco, CA*, 1996.

[5] A. O. Artero and M. C. F. de Oliveira. Viz3d: Effective exploratory visualization of large multidimensional data sets. In *The 17th Symposium on Computer Graphics and Image Processing 2004, Brazilian*, pages 340–347, 2004.

[6] S. K. Card, J. D. Mackinlay, and B. Schneiderman. *Readings in information visualization: using vision to think.* Morgan Kaufmann, 1999.

[7] K. Daniels, G. Grinstein, A. Russell, and M. Glidden. Properties of normalized radial visualizations. *Information Visualization*, 11(4):273–300, 2012.

[8] L. di Caro, V. Frias-martinez, and E. Frias-martinez. Analyzing the role of dimension arrangement for data visualization in radviz. In *Advances in Knowledge Discovery and Data Mining*, pages 125–132, 2010.

[9] J. Havrda and F. Charvát. Quantification method of classification processes: Concept of structural α-entropy. *Kybernetika*, 3(1):30–35, 1967.

[10] 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 1997*, pages 437–441, 1997.

[11] 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*, pages 9–16, 1999.

[12] J. Ian. *Principal Component Analysis.* Wiley Online Library, 2005.

[13] A. Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1(2):69–91, 1985.

[14] E. Kandogan. Star coordinates: A multi-dimensional visualization technique with uniform treatment of dimensions. In *Proceedings of the IEEE Information Visualization Symposium 2000*, volume 650, pages 4–8, 2000.

[15] D. A. Keim, M. Ankerst, and H.-P. Kriegel. Recursive pattern: A technique for visualizing very large amounts of data. In *Proceedings of the 6th Conference on Visualization'95*, pages 279–286, 1995.

[16] G. Leban, B. Zupan, G. Vidmar, and I. Bratko. Vizrank: Data visualization guided by machine learning. In *Data Mining and Knowledge Discovery 13*, pages 119–136, 2006.

[17] D. J. Lehmann and H. Theisel. Orthographic star coordinates. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2615–2624, 2013.

[18] X. Li, K. Zhang, and T. Jiang. Minimum entropy clustering and applications to gene expression analysis. In *Computational Systems Bioinformatics Conference, (CSB 2004)*, pages 142–151, 2004.

[19] J. McCarthy, K. Marx, P. Hoffman, A. 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*, 1020(1):239–262, 2004.

[20] M. Rubio-Sanchez and A. Sanchez. Axis calibration for improving data attribute estimation in star coordinates plots. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2013–2022, Dec. 2014.

[21] M. Sips, B. Neubert, J. P. Lewis, and P. Hanrahan. Selecting good views of high-dimensional data using class consistency. *Computer Graphics Forum*, 28(3):831–838, 2009.

[22] T. Van Long and L. Linsen. Visualizing high density clusters in multidimensional data using optimized star coordinates. *Computational Statistics*, 26(4):655–678, 2011.