REGULAR PAPER

Fangfang Zhou · Bing Bai · Yitao Wu · Minghui Chen · Zengsheng Zhong · Rongchen Zhu · Yi Chen ·
Ying Zhao

FuzzyRadar: visualization for understanding fuzzy clusters

Abstract Fuzzy clustering assigns a membership degree (MD) on a datum to a cluster, which reflects real-world clustering scenarios but increases the complexity of understanding fuzzy clusters. Many studies have demonstrated that multidimensional visualization techniques are beneficial to fuzzy clusters analysis. However, empirically, no single existing visualization technique can support most analytical tasks featured by fuzzy clustering. This work proposes a new visualization called FuzzyRadar for understanding fuzzy clusters. Its basic idea is to combine the advantages of radial coordinate visualization (Radviz), which specializes in data-oriented analytical tasks, and parallel coordinate plot (PCP), which performs well in cluster-oriented analytical tasks. First, we adopt a compact and compounded layout to integrate Radviz and PCP into one visualization view. Then, we introduce a strip-edge-bundling method to reduce the visual cluster caused by PCP polylines and a histogram embedding method to facilitate the recognition of MD distribution. We also provide a group of additional visual encodings and a set of lightweight interactions. Finally, we use a case study to demonstrate the usability of FuzzyRadar and conduct a controlled quantitative evaluation to compare the performance of FuzzyRadar, Radviz, PCP, and scatterplot matrix. Result shows that FuzzyRadar supports all the seven examined analytical tasks well and presents a significant capability improvement compared with Radviz and PCP.

Keywords Visualization · Visual analysis · Fuzzy clustering · Radviz · Parallel coordinate plot

1 Introduction

Traditional hard clustering assorts objects into a certain cluster without considering uncertainty; however, this is not always realistic. Soft clustering (Klawonn et al. 2003), known as fuzzy clustering, provides a natural approach for handling the uncertainty in assorting objects, which accepts that clusters in data are usually not completely separated; thus, a membership degree (MD) between 0 and 1 for each cluster is

F. Zhou · Y. Wu · M. Chen · Z. Zhong · R. Zhu · Y. Zhao (✉)
School of Computer Science and Engineer, Central South University, Changsha, China
E-mail: zhaoying@csu.edu.cn

B. Bai
School of Automation, Central South University, Changsha, China

Y. Chen
Beijing Key Laboratory of Big Data Technology for Food Safety, Beijing Technology and Business University, Beijing, China

assigned to every datum. Fuzzy clustering has been widely accepted as a preferable solution due to its reflection of real-world clustering scenarios.

Fuzzy clusters generated by fuzzy clustering are commonly expressed as a MD matrix, in which rows and columns describe data items and clusters, respectively, and a cell indicates the MD of a certain datum to the corresponding cluster. When the MD matrix contains numerous data items and a plurality of clusters, the data would become complex multidimensional or even high dimensional. This is a tough challenge faced by analysts in attempting to gain insights into fuzzy clusters. As visualization has become an important technique in various domains for understanding complex data (Zhou et al. 2017, 2019a, b; Zhao et al. 2018; Wu et al. 2016; Liu et al. 2017; Bi et al. 2019), many multidimensional visualization methods have been introduced to analyze the MD matrix in an interpretable and interactive manner (Feil et al. 2007; Sharko and Grinstein 2009; Zhou et al. 2017; Lin et al. 2015).

Empirically, one visualization technique performs well only on a particular analysis task (Dimara et al. 2018). A state-of-the-art evaluation study (Zhao et al. 2019) recently confirmed this point. The study concluded seven analytical tasks featured by fuzzy clusters analysis and systematically evaluated the performance of four multidimensional visualization techniques, namely parallel coordinate plot (PCP), scatterplot matrix (SPM), principal component analysis (PCA), and radial coordinate visualization (Radviz), in analyzing fuzzy clusters. The evaluation results showed that no single visualization technique had remarkable capability to support all the tasks well. The results also showed that Radviz obtained the best overall performance in data-oriented tasks, which mainly benefits from its radial spring-based projection mechanism. PCP outperformed the other three techniques in cluster-oriented tasks due to its vertical axes that represent clusters.

On the basis of the evaluation result (Zhao et al. 2019), we bring out a new idea that combines the advantages of Radviz and PCP and designs an improved multidimensional visualization technique. This technique is expected to support most analytical tasks of understanding fuzzy clusters well. However, many design challenges exist in practice. First, integrating the two techniques directly is difficult. Radviz presents a compact and radial layout, whereas PCP has a loose and rectangular layout. Radviz and PCP are entirely different in visual encodings of data presentation. Second, reducing the visual clutter caused by polylines in PCP is crucial. Third, Radviz and PCP are inadequate in presenting the information of dominant cluster and MD distribution. We need to provide new visual encodings to address this issue. In addition, interactions are needed to help users conduct smooth analysis by using Radviz and PCP collaboratively.

In this study, a new visualization called FuzzyRadar is proposed to help users understand fuzzy clusters. We adopt a compact and compounded layout to integrate Radviz and PCP into one visualization view. This layout keeps the original visual encodings of Radviz and PCP as far as possible without additional visual clutter. We introduce a strip-edge-bundling method that replaces each group of bundled PCP polylines with a polygonal strip, which is beneficial not only to reduce the visual cluster caused by PCP polylines but also to facilitate dominant cluster identification. We also introduce a histogram embedding method to explicitly present the MD distribution of all data items on a cluster. Moreover, we design a series of additional visual encodings to enhance the capability of FuzzyRadar in identifying data stability and dominant cluster. We also provide a set of lightweight interactions that enable users to conduct smooth interactive analysis.

To evaluate the capability of FuzzyRadar, seven typical analytical tasks of understanding fuzzy clusters are formulated, and four real-world data sets are collected. First, we use a case study to demonstrate the usability of FuzzyRadar. Then, we conduct a controlled quantitative experiment to compare the performance of FuzzyRadar, Radviz, PCP, and SPM. Fifteen undergraduate students are recruited as experiment subjects. Questionnaires are designed to collect two objective metrics (namely, accuracy and time) and a subjective metric (namely, satisfaction) as experiment results. The experiment results confirm that FuzzyRadar generally outperforms Radviz, PCP, and SPM in completing all the seven analytical tasks. The results also suggest that FuzzyRadar presents a significant capability improvement compared with each of Radviz and PCP. Finally, we discuss the limitations of this work and future work directions.

2 Related work

2.1 Multidimensional data visualization and evaluation

Many visualization techniques exist to visualize multidimensional data (Xia et al. 2018a, b; Ma et al. 2019; Xie et al. 2014). We classify them into two main categories, namely, lossy and lossless techniques. Lossy

techniques commonly utilize linear or nonlinear dimensionality reductions (DR) to map multidimensional data points into 2D/3D observable plane. PCA and linear discriminant analysis are the most representative linear methods, whereas Radviz and t-SNE are typical nonlinear methods. The common shortcoming of lossy techniques is that the process of DR will inevitably cause the loss of information in original space. Meanwhile, lossless techniques present all dimensions of a data set simultaneously to avoid such information loss. PCP (Inselberg and Dimsdale 1987) and SPM (Long 2014) are well known in this category. Nevertheless, as the number of dimensions increases, visualizing many dimensions in a limited screen space often produces serious visual clutter.

Visualizing multidimensional data on an observable plane is an inherently ill-posed problem; thus, all methods have drawbacks. Researchers have gradually realized the importance of comparing and evaluating different techniques (Manuel et al. 2016; Etemadpour et al. 2015; Rzeźniczak 2013). In terms of the lossy techniques, Sedlmair et al. (2013) explored the performance of four frequently used DR methods on class separability. For the lossless techniques, Holten and Van Wijk (2010) evaluated the time and correctness performances of nine PCP variations for cluster identification. With respect to application scenarios, Rajanen (Marghescu 2007) evaluated different multidimensional data visualization techniques in solving the problem of financial competitor benchmarking.

This work proposes a new visualization that combines the advantages of PCP and Radviz for understanding fuzzy clusters. This work also presents an experiment to evaluate the performance of the proposed visualization compared with Radviz, PCP, and SPM.

2.2 Fuzzy clusters visualization

Fuzzy c-means (FCM) clustering algorithm (Zadeh 1965; Ruspini 1969; Dunn 1973) may be the most widely known fuzzy clustering algorithm. Given that results of fuzzy clustering are typical multidimensional data, researchers have proposed various methods to visualize fuzzy-clustered data (Klawonn et al. 2003). Lossy techniques are frequently used to visualize fuzzy clusters. De Oliveira and Pedrycz (2007) and Hoppner and Klawonn (2006) utilized PCA and multidimensional scaling (MDS) to present a fuzzy clustering result on a 2D plane. Abonyi and Babuska (2004) proposed a modified fuzzy Sammon mapping (FSM) to visually analyze the associations between data points and clusters. Rueda and Zhang (2006) proposed a novel method that maps data points into an irregular hyper-tetrahedron to reflect inter-cluster relationships in a spatial manner. Sharko and Grinstein (2009) and Zhou et al. (2017) found that Radviz is advantageous in visualizing fuzzy clusters because the projecting position of each data point in Radviz can directly indicate its memberships belonging to all clusters. To avoid the information loss of lossy techniques, some lossless approaches have been introduced to present the complete information of fuzzy-clustered data. Berthold and Hall (2003) utilized PCP to visualize fuzzy clusters. On the basis of the design of SPM, Lin et al. (2015) and Cao et al. (2015) presented a novel technique called UnTangle Map that weaves an interactive mesh of triangle-style scatterplots for the analysis of fuzzy clusters.

Moreover, a few studies have evaluated the advantages and disadvantages of the aforementioned methods. For example, Abonyi and Babuska (2004) compared the cluster validity performance of FSM and PCA on fuzzy clusters. Lin et al. (2015) conducted a brief evaluation to illustrate the advantages of UnTangle Map over PCP, SPM, and PCA. Zhao et al. (2019) conducted a state-of-the-art study to evaluate multidimensional visualization techniques in analyzing fuzzy clusters. PCP, SPM, PCA, and Radviz were evaluated. The study confirmed that no single existing visualization technique had remarkable capability to support all the tasks featured by fuzzy clusters analysis well. The study also provided instructive guidance for technique selection. Inspired by this study, we design a compounded visualization to help users analyze fuzzy clusters.

2.3 Uncertainty visualization

As an integral part of data, uncertainty is generally used to describe the incorrectness, incompleteness, and ambiguity of data. Visualization techniques (Chen et al. 2018a; b; Shi et al. 2016, 2018) can incorporate such uncertainty into visual representation, thereby communicating uncertain information with users instead of neglecting them. Considerable effort has been made to design uncertainty visualizations that help users make better decisions (Zuk and Carpendale 2006; Epp and Bull 2015; Wu et al. 2012; Chen et al. 2015). In addition, researchers have evaluated whether and how users perceive and interpret meaningful uncertainty information from visual representation. For example, Sanyal et al. (2009) constructed a user study to

evaluate the effectiveness of four commonly used uncertainty visualization techniques for 1D and 2D data sets. Ferreira et al. (2014) compared six different approaches of encoding temporal uncertainty in terms of error and completion time. The results of fuzzy clustering are a type of multidimensional data that contain classification uncertainty. This work proposes a new visualization to help users understand the uncertainty information in fuzzy clusters.

3 Scenario and task

Various application scenarios need understanding of fuzzy clusters. For example, online music providers would like to know the music preferences of their customers to provide highly personalized services (Zhao et al. 2019). Generally, customers have different music preferences. Some customers only like pop music, whereas some like pop and jazz music to different degrees. Music preferences are a type of fuzzy clusters where customers and music types are data items and clusters, respectively. Music providers would want to determine the answers of some specific queries, such as which types of music are popular, whether the preferences of a certain customer are clear, and which groups of customers have similar preferences.

After a review of the related literature, we establish seven typical analytical tasks of understanding fuzzy clusters. These tasks are divided into two categories, namely data-oriented (T1–T3) and cluster-oriented (T4–T7) tasks. Among the data-oriented tasks, T1 and T2 are about single item, whereas T3 is about multiple items. Among the cluster-oriented tasks, T4 and T5 are single-cluster oriented, whereas T6 and T7 are multi-cluster oriented. The tasks are detailed as follows.

T1. Membership information of a data item What are the maximum and minimum MDs of a given data item? In the aforementioned scenario, the maximum and minimum MDs of a data item represent the most and least favorite music types of a certain customer, respectively.

T2. Stability of a data item Is a data item stable? This information indicates whether a data item has a dominant MD. In the scenario, if a customer has a dominant MD (i.e., MD $\in$ [0.8, 1]) to pop music, this customer is stable because he/she has a dominant preference on pop music.

T3. Stability of a data group Are more stable data items than unstable data items presented in a given data group? In the scenario, if the MDs of a data group to pop music are largely distributed between 0.8 and 1.0, the group of customers are stable.

T4. Membership information of a single cluster What are the maximum and minimum MDs of a given cluster? For pop music, the maximum and minimum MDs represent the maximum and minimum preference degrees among customers who prefer pop music, respectively.

T5. Dominant cluster Does a dominant cluster exist in a fuzzy clustering result? On the basis of the maximum MD principle, if the number of data items partitioned into a cluster is significantly greater than that of other clusters, then the cluster is dominant. The maximum MD principle is a strategy that is commonly used to convert soft clustering result into hard clustering result. Particularly, when classifying a data item that belongs to multiple clusters into one cluster, this principle states that the data item should be classified into the cluster that corresponds to the maximum MD of the data item.

T6. Similarities between clusters Given multiple clusters, do they have similar membership distributions? Through this task, we can identify similarity between two clusters.

T7. Correlations between clusters Given multiple clusters, do they have positive correlations? If the MDs of most data items belonging to two clusters increase or decrease simultaneously, then the two clusters have a positive correlation. In the scenario, if customers who like pop music tend to like jazz as well, these two music types have a positive correlation.

4 Design challenges

On the basis of the state-of-the-art evaluation study (Zhao et al. 2019), we can acquire four insightful and experienced guidelines to improve the capability of multidimensional visualizations in understanding fuzzy clusters. (1) Suitable data projection mechanism allows users to quickly determine the stability of data items. (2) Axes corresponding to clusters facilitate cluster-oriented information recognition. (3) Low visual clutter is crucial to reducing recognitive burden. (4) Interactions play an important role in solving analytical tasks.

On the basis of the four guidelines, we attempt to design a new multidimensional visualization technique for understanding fuzzy clusters. The evaluation study (Zhao et al. 2019) has proved that Radviz is the best technique for data-oriented tasks (T1–T3), whereas PCP is the best one for cluster-oriented tasks (T4–T7); thus, our basic idea is to combine the advantages of Radviz and PCP. The combination of the two techniques is expected to preferably support all tasks. However, to achieve this goal, the following design challenges still exist:

H1: How to design a suitable combined layout of Radviz and PCP The visual encodings of Radviz and PCP are entirely different. Radviz presents a compact and radial layout, whereas PCP has a loose and rectangular layout. Data items are represented by points in Radviz but by polylines in PCP. Therefore, combining the two visualization techniques efficiently is difficult.

H2: How to present an explicit MD distribution When performing T5 and T6, users need to recognize the MD distribution of all data items on each cluster. However, such MD distribution information in PCP is presented implicitly. Therefore, users must view all MDs on an axis to estimate the relevant MD distribution. This process is a visual burden on users and is time-consuming and error-prone.

H3: How to reduce the visual clutter caused by polylines in PCP When numerous data items exist, an excessive overlapping of polylines will cause severe visual clutter in PCP. The visual clutter limits analysis efficiency of PCP in cluster-oriented tasks (T4 and T7).

H4: How to enhance the visual presentation of data item stability and dominant cluster T2 and T5 are the most difficult tasks because data item stability and dominant cluster are abstract concepts, and no relevant visual presentation is directly provided by Radviz and PCP. Therefore, users have to search for multifarious information to solve the tasks.

H5: Provide lightweight and simple interactions We must provide appropriate interactions to help users perform a smooth analysis process by using the two visualization techniques simultaneously. Given that users might lack experience of visual analysis, the interactions must be lightweight and easy to use.

5 Visualization and interaction

5.1 Basic layout

There are two basic ideas to address H1, namely juxtaposed and compounded. Two layout design alternatives can be derived from the juxtaposed idea, as shown in Fig. 1a, b. Radviz and PCP are arranged side by side with two types of space partition methods, that is, (1) Radviz and PCP share the same horizontal size (Fig. 1a), which limits PCP because it generally needs a large horizontal space to layout its axes; and (2) Radviz and PCP have the same vertical size (Fig. 1b), which weakens Radviz because the entire display space of Radviz will be small if PCP uses a large horizontal space to present a number of axes. In addition, the juxtaposed idea must use two independent visualization views, which will occupy considerable screen space. Accordingly, we exclude the juxtaposed idea.

The compounded idea also has two layout design alternatives, as shown in Fig. 1c, d. Radviz and PCP are combined into one compact and compounded visualization view with two types of fusion styles, namely inner and outer fusion. In the inner fusion style (Fig. 1c), the dimension arcs of Radviz are used as the PCP axes, each of which is marked with an MD interval of [0, 1], and the MD values are arranged in a clockwise manner. The positions of data points inside the Radviz circle are still determined by the Radviz projection mechanism. Each data point is linked to the corresponding MD positions on all dimension arcs by multiple lines. However, the data points and formed lines are all presented inside the Radviz circle, thereby causing severe visual clutter. We finally decide to use the outer fusion style of the compounded layout (Fig. 1d). The PCP axes are placed at the corresponding dimensional anchors of Radviz and extend outward the Radviz circle. This layout is compact and harmonious. It keeps the original encodings of Radviz and PCP as far as possible and introduces no additional visual clutter.

An initial visualization result is shown in Fig. 1e. The Radviz circle is equally divided into multiple colored arcs to present clusters labeled with *C*1, *C*2, …, *Cm*. The clusters' order is determined based on the similarity between clusters (Peng et al. 2004; Yang et al. 2003). Each data item is represented by a data point in the Radviz circle, and the color of a data point indicates the cluster that the data item belongs to on the basis of the maximum MD principle. PCP is placed outside the Radviz circle. The starting positions of the PCP axes are located at the central positions of the corresponding arcs. The PCP polylines are converted into curves, and the colors of the curves are consistent with those of the corresponding data points.

5.2 Embedded MD histograms

Histogram is the most common manner to explicitly present the MD distribution of all data items on a cluster (H2). Generally, two design alternatives embed a histogram into a PCP axis, namely off-axis and in-axis. In the off-axis style, a histogram is embedded outside an axis; however, in this style, polylines will overlap with the histogram bars and cause additional visual clutter, as shown in Fig. 2b. In the in-axis style, each axis is duplicated into a pair; a histogram is located between the axis pair (Fig. 2c), and polylines are drawn between duplicated axis pairs. This method is additional-clutter-free, and users can clearly recognize MD distributions from histograms.

We use the in-axis style, as shown in Fig. 2a. The specific steps of drawing histograms are as follows.

(a) Divide the MD interval, and determine the values of subintervals. First, divide the MD interval [0, 1] into 10 subintervals, [0, 0.1], [0.1, 0.2], …, [0.9, 1]. Then, for each subinterval on each cluster, calculate the value of the subinterval, that is, the number of data items falling into it. As a result, the subinterval value matrix N is obtained, as shown as follows:

$$N = \begin{pmatrix} n_{11} & n_{12} & \dots & n_{110} \\ n_{21} & n_{22} & \dots & n_{210} \\ \dots & \dots & \dots & \dots \\ n_{c1} & n_{c2} & \dots & n_{c10} \end{pmatrix}$$

where c is the number of clusters, and n_{ij} ($1 \le i \le c$, $1 \le j \le 10$) is the number of data items falling into the subinterval [$0.1 \times (j-1)$, $0.1 \times j$] on cluster i.

(b) Determine the angle of each axis pair, namely the angle that each axis pair occupies on the circumference of Radviz circle. Set the total angle of axis pairs to a fixed value, that is, $360° \times \alpha$, where α is an adjustable parameter. Then, the angle of each axis pair is determined as $360° \times \alpha/c$, where c is the number of axis pairs (clusters).

(c) Determine the angles of histogram bars. Since the histograms are placed on the Radviz circle, the histogram bars are curved. Therefore, we use angles to measure the lengths of bars. In this case, the angle of a bar represents the corresponding subinterval value, and the maximum angle A_{max} is equal to the angle of each axis pair, $360° \times \alpha/c$. First, find the maximum subinterval value $n = \max(N)$ and assign the maximum angle A_{max} to its bar. Then, for each subinterval, calculate the value ratio n_{ij}/n and assign $A_{max} \times n_{ij}/n$ to the angle of its bar.
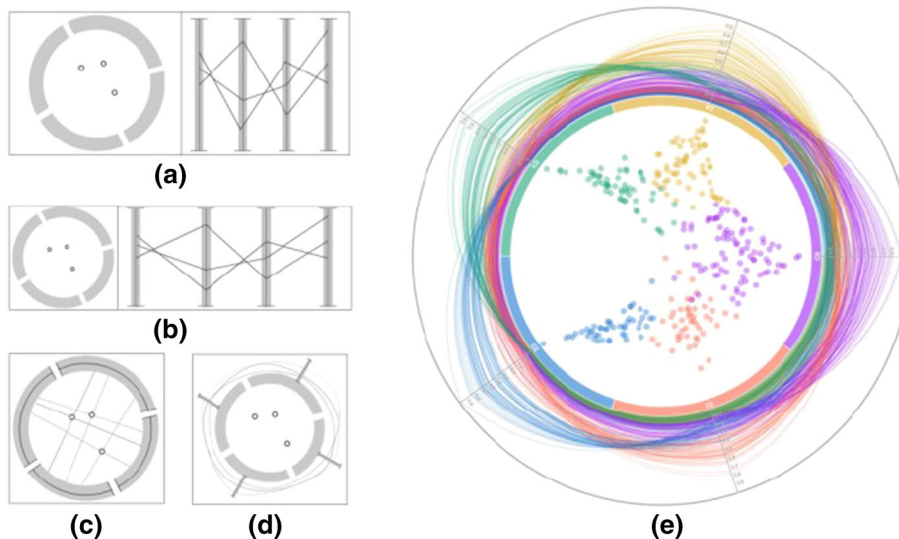
Fig. 1 Design alternatives of the basic layout for combining Radviz and PCP. **a** Juxtaposed layout with the same horizontal size; **b** juxtaposed layout with the same vertical size; **c** compounded layout with an inner fusion style; **d** compounded layout with an outer fusion style; **e** initial visualization result with the final basic layout design
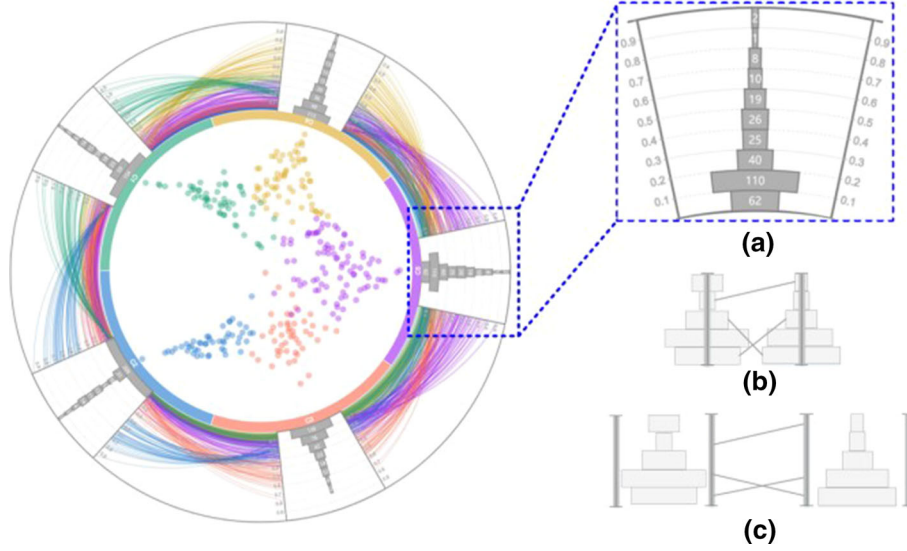
Fig. 2 **a** Visualization design of embedded histograms and two design alternatives: **b** off-axis style and **c** in-axis style

(d)   Add the bars. For each subinterval, place the bar at the center of the axis pair, and mark the number of data items on the bar.

5.3 Strip-edge-bundling

A common manner to reduce the visual clutter caused by polylines in FuzzyRadar (H3) is to bundle the polylines with the same color between axes. The key of edge-bundling method is how to set the control points of Bezier curves. In our case, polylines are arc-shaped because PCP is placed outside the Radviz circle. Therefore, we cannot directly use the existing edge-bundling method (Palmas et al. 2014). We therefore propose a specific strip-edge-bundling method for FuzzyRadar. A schematic illustration is shown in Fig. 3, including the four steps.

(1)   Add virtual axes A' and B' for adjacent coordinate axes A and B, respectively. The horizontal distance between each virtual axis and its original axis is set to 10% of the distance between axes A and B.

(2)   Group polylines on original axes by colors. As shown in Fig. 3a, three groups exist on axis A, namely red, blue, and green groups. Taking the red group as an example, map the MD range R (y^-, y^0, y^+) of the red group of data items to R' (y'^-, y'^0, y'^+) on virtual axis A', where y^- and y^+ are the minimum and maximum MDs, respectively, and y^0 is the mean MD on axis A. Let $y'^0 = y^0$, $y'^- = y'^{0-} - 0.5 \times W \times \beta$, and $y'^+ = y'^0 + 0.5 \times W \times \beta$, where W is the number of data items belonging to the red group, and β is a parameter for controlling the vertical distance of the red group on virtual axis A'.

(3)   Draw the Bezier curve of a certain data item X in the red group. Virtual axes A' and B' divide the entire curve into three segments. Use a quadratic Bezier curve for each of the first and third segments. As shown in Fig. 3a, the red squares in the first and third segments are the control points of the quadratic Bezier curves. For the second segment, a cubic Bezier curve is used. First, draw the center line P of axes A' and B'. Second, draw the center line P' between axes A' and P. Third, draw a tangent line at X_A, and its intersection point with P' is the first control point. Fourth, draw a tangent line at X_B and find its intersection point with line P. Subsequently, take the midpoint between the intersection and the vertex of line P as the second control point. As shown in Fig. 3a, the red squares in the second segment are the control points of the cubic Bezier curve.

(4)   Polylines by groups are bundled closely and stick together after the former three steps, thereby reducing the visual clutter. However, recognizing the number of polylines in a bundle for users is difficult; therefore, users cannot compare the number of data items between groups. To address this problem, we replace each Bezier curve bundle with a polygonal strip. As shown in Fig. 4, the red
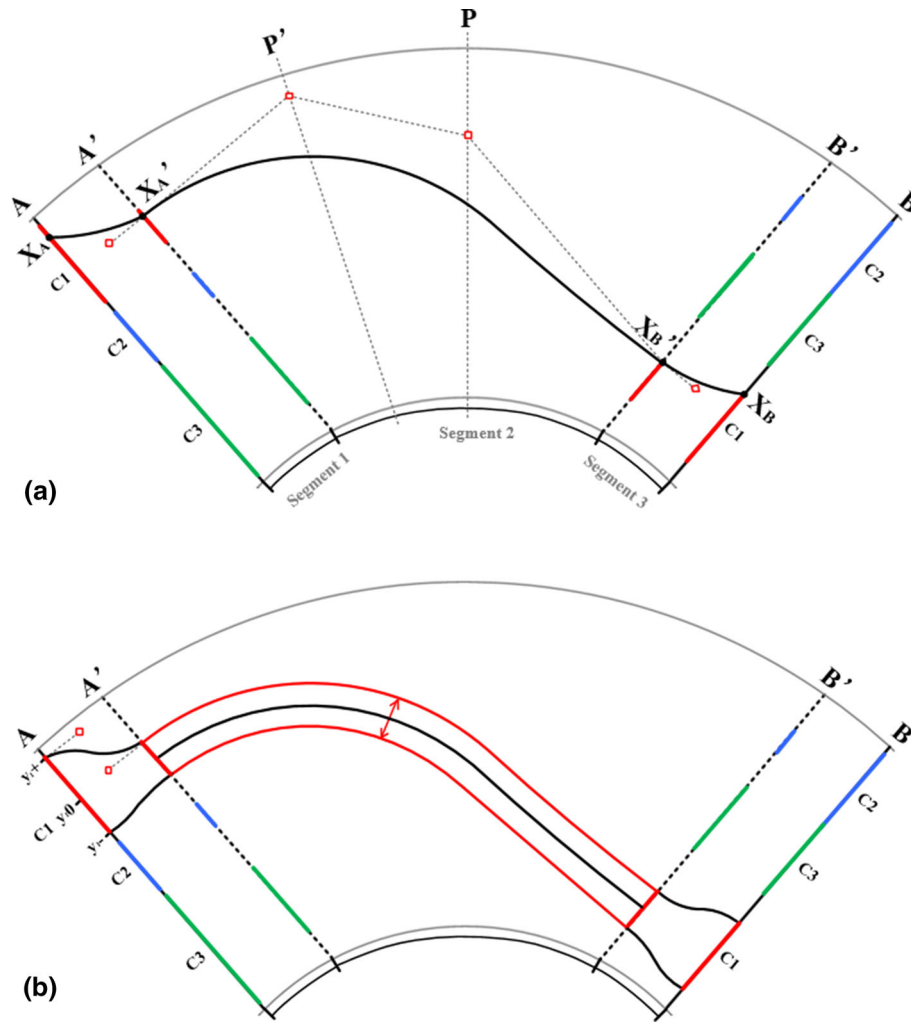
Fig. 3 Illustration of our **a** edge-bundling and **b** strip-edge-bundling methods

Bezier curve bundle is replaced with a single strip, the width of which represents the number of bundled curves.

A visualization result after strip-edge-bundling is shown in Fig. 4. In comparison with Fig. 2b, the visual clutter is largely reduced. We can find that the purple strip has the maximum width among all colored strips. That is, the corresponding purple cluster has more data items than the other clusters according to the principle of maximum MD. This observation is beneficial for users to identify the dominant cluster (T5; See Sect. 6.1).

5.4 Encodings of data stability and dominant cluster

We use double or triple visual encodings to help users perceive data stability or the dominant cluster (H4). For data stability, we use double encodings. The first encoding is the positions of data points contributing to Radviz's projection mechanism. If a data point is pulled near a Radviz anchor, then it is stable (Fig. 4-①). By contrast, if a data point is located at the circle center or the area between two anchors, then it is unstable (Fig. 4-②, -③). The second encoding is color. We add a ring outside each data point. The darker the ring, the more stable the data point is (Fig. 4-①).

Dominant cluster is determined by cluster sizes, namely the number of data items belonging to a cluster according to the maximum MD principle. We use triple encodings to present cluster sizes. The first encoding is color. The color of a data point is consistent with the color of its relevant cluster. Thus, users can judge the
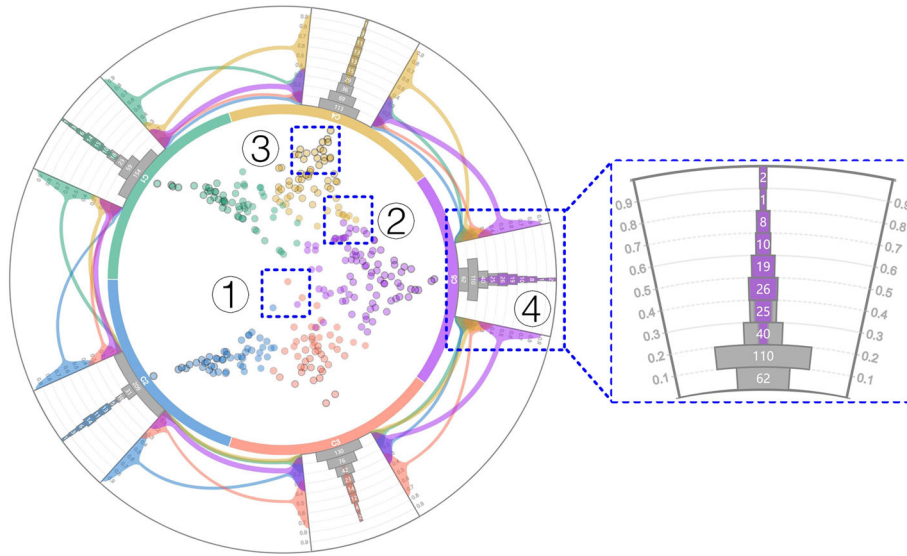
Fig. 4 Visual encodings of data stability and colored histogram bars

size of a cluster by observing the number of data points with the same color. The second encoding is the widths of colored strips (See Sect. 5.3). The third encoding is the colored areas in histograms. As shown in Fig. 4-④, in each histogram bar, the colored area indicates the number of data items that belong to the relevant cluster with the MDs falling in the subinterval.
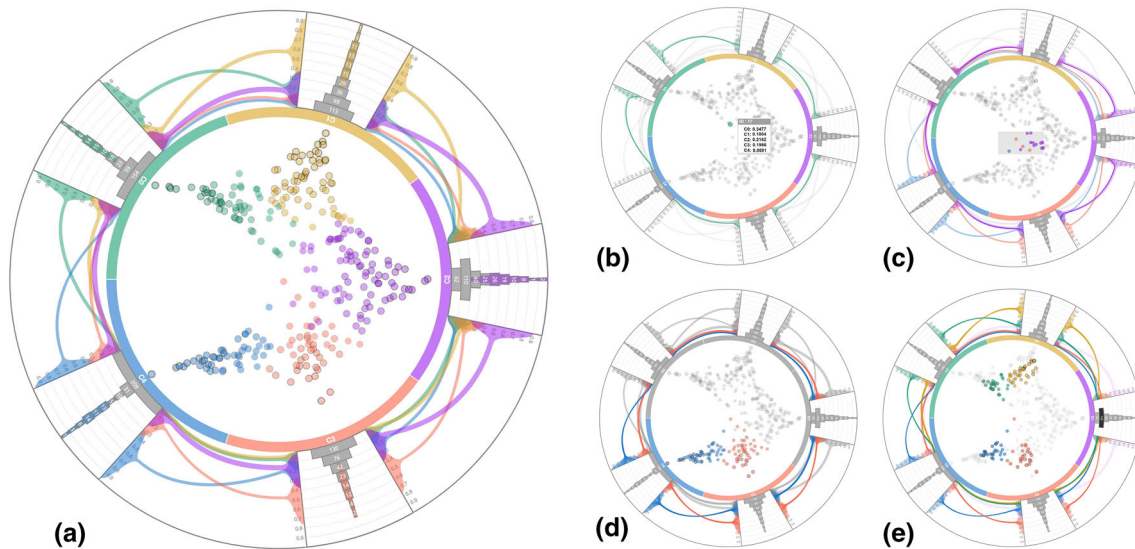
5.5 Interactions

We provide four lightweight interactions in FuzzyRadar to help users conduct smooth analysis by using Radviz and PCP collaboratively (H5). The four interactions are detailed as follows.

1. Information notification. When the mouse is hovering on a visual element, the detailed information will be displayed in a pop-up tip and other relevant visual elements will be highlighted simultaneously. As shown in Fig. 5b, the MD information of a selected data point is shown in a pop-up tip, and the relevant polyline is highlighted at the same time.
2. Area selection. Users can brush a rectangle to select a group of data points within the Radviz circle, and the corresponding polylines will be highlighted in PCP, as shown in Fig. 5c.
3. Cluster selection. Users can select a cluster or multiple clusters by clicking cluster arcs, and the relevant data points and strips will then be highlighted, as shown in Fig. 5d.
4. MD subinterval selection. When the mouse is hovering on a bar in a histogram, the relevant data points and polylines will be highlighted, as shown in Fig. 5e.

6 Evaluation

6.1 Case study

For the case analysis, this section uses the Heart_Disease data (Asuncion and Newman 2007) collected at the Cleveland Clinic Foundation in the UCI Machine Learning Repository. The Heart_Disease data contain 303 data items and 14 distinct dimensions. These dimensions represent the patients' age, gender, angina, resting blood pressure, and other information. The last dimension is the categorical attribute of the Heart_Disease data, which are marked C0–C4, which correspond to five severity levels of heart disease. Among them, C0 represents no heart disease, and C4 represents the severest heart disease. We use the classic FCM algorithm to obtain the fuzzy clustering result of the data. Figure 5a shows the visualization result of the fuzzy clustering result by using FuzzyRadar.

Fig. 5 Interactions and case study. **a** Visualization result of fuzzy clusters using the Heart_Disease data. **b** Select an unstable data point to show relevant MD information. **c** Select a group of data points. **d** Select clusters C3 and C4. **e** Select the data points that belongs to C4 with 0.1–0.2 MD values

We observe the overall stability of data items (T3). The majority of data points within the circle have relatively dark border rings and are located near the dimension anchors. This observation indicates that the number of stable data items in the fuzzy clustering result is far more than that of unstable data items. Generally, a stable data point reflects that the corresponding patient has a certain diagnosis, whereas an unstable data item represents that the diagnosis of the corresponding patient is under high uncertainty. Subsequently, we select one unstable data point for analysis (T1 and T2). As depicted in Fig. 5b, the selected data point (No. 17) is located near the center of the circle, thereby indicating its high instability. According to the principle of maximum MD, the No. 17 data point belongs to the green C0 cluster (no heart disease). However, the data point presents relatively high MDs to several clusters (C0: 0.3477; C1: 0.1804; C2: 0.2152; and C3: 0.1986). Consequently, the patient still has a high risk of suffering from heart disease and thus requires regular and careful checkups.

We analyze the similarities between clusters (T4 and T6) using the histograms outside the circle. As depicted in Fig. 5a, clusters C0, C1, and C3 have similar MD distributions, presenting a slow downward trend from low to high MD (extending from the circle to the outside). Meanwhile, C4's MD distribution shows a steep decline trend because large quantities of data items present low MD values (≤ 0.2) to C4; hence, patients generally have low risks of suffering from the severest heart disease. Moreover, only C2 out of all clusters does not show a monotonous downward trend in its MD distribution, which arouses our interest. As shown in the histograms, the histogram of C2 apparently has a larger colored (purple) area than those of other clusters. As shown in the colored strips outside the circle, the purple strip of C2 is the widest among all strips. These two observations reflect that C2 may be a dominant cluster. According to the principle of maximum MD, the numbers of data items separately belonging to clusters C0–C4 are 56, 59, 87, 46, and 49, respectively. C2 accounts for 87, which is half more than the other clusters. Consequently, C2 is a dominant cluster; thus, patients generally have high risks of suffering from the relatively mild heart disease.

We analyze the correlations between two adjacent clusters using two cases. The first case uses clusters C1 (yellow) and C2 (purple). We divide the colored strips between the histograms of the two clusters into two groups, as shown in Fig. 5a. That is, all data items are divided into two groups. In the first group, the yellow C1 strip and the purple C2 strip present a crisscross shape, which indicates that the first group of data items show a negative correlation between their MD values to C1 and to C2. In the second group, the strips of C0, C3, and C4 present a parallel shape, which indicates that the second group of data items show a positive correlation between their MD values to C1 and to C2. Two groups of colored strips are nearly of equal aggregated width; thus, the number of data items with the positive correlation is almost equal to that of data items with the negative correlation. Consequently, no significant positive correlation is found between clusters C1 and C2. The second case uses clusters C3 (orange) and C4 (blue), as shown in Fig. 5d. We

divide the colored strips between the histograms of the two clusters into two groups. In the first group, the orange C3 strip and the blue C2 strip present a crisscross shape. In the second group, the strips of C0, C1, and C2 present a parallel shape. The comparison of the aggregated widths of the two groups of strips indicates that the data items with the positive correlation are apparently considerably more than those with the negative correlation. Consequently, a positive correlation exists between clusters C3 and C4.

6.2 User study

We conducted a user study to verify the performance of FuzzyRadar. We selected three classic multidimensional data visualization methods (namely PCP, SPM, and Radviz) for comparative analysis. We used the fuzzy clustering results of Iris (for training) (Asuncion and Newman 2007), Synthetic (Zhou et al. 2016), Heart_Disease (Asuncion and Newman 2007), and Concrete (Asuncion and Newman 2007) data sets as experimental data. We recruited 15 voluntary college students. Their age ranges from 21 to 27 (average 24). They are all graduate students affiliated with the school of computer science and engineering in a university.

On the basis of the evaluation study (Zhao et al. 2019), we designed questionnaires, implemented an experiment system, and conducted a controlled experiment. The experiment comprised three phrases. (1) First was the tutorial phrase. An experimental instructor explained the purpose and tasks of the experiment at the beginning and helped the volunteers become familiar with the procedure, experiment system, and questionnaires. Then, the volunteers were provided with the Iris data set to perform a pre-experiment, in which the volunteers must answer several analytical questions related to the Iris data set. (2) Second was the formal study phrase. Each volunteer was required to analyze a randomly assigned data set using all the four visualization techniques separately. A combination of a technique and a data set had a specifically defined questionnaire, which contains 10–15 questions that cover the seven analytical tasks and are featured by the relevant data set. After answering each question, the volunteers must rate their satisfaction of using this technique in solving the question on a seven-point Likert scale ranging from 1 (strongly dissatisfied) to 7 (strongly satisfied). The data sets and the four visualization techniques appeared in order of the Latin square sorting. (3) Third was the interview phrase, in which we had a brief interview with the volunteers after the formal study phrase.

We fully recorded two objective metrics and a subjective metric as experiment results. Accuracy and time of using each technique in completing each question were collected as the two objective metrics. Satisfaction ratings, as the subjective metric, were recorded to disclose the volunteers' technique preferences for solving each question. Processing the experiment results included normality tests, testing of significant differences, and pairwise comparisons. We initially used the Shapiro–Wilk test to examine the normality and found that all metrics and their results did not follow normal distributions. Thus, we used a nonparametric Friedman test and a Tukey's HSD test to examine whether the four techniques have significant differences in the three metrics. All the tests were performed under the standard significance level of $p = 0.05$.

The result of mean accuracy in solving each task with the four visualization techniques is shown in Fig. 6a. All visualization techniques perform well in terms of accuracy for the relatively simple analytical tasks T1 and T4. FuzzyRadar obtains the highest accuracy for tasks T2, T3, T5, and T6. Especially for T6, FuzzyRadar remarkably outperforms the other three techniques. This result reflects that the embedded histograms of displaying MD distribution information can facilitate the analysis of cluster similarities. For T7, FuzzyRadar outperforms Radviz and PCP in terms of accuracy, which indicates that our strip-edge-bundling method can improve the ability of cluster correlation analysis of Radviz and PCP. However, FuzzyRadar still underperforms SPM because SPM is always outstanding in correlation analysis (Zhao et al. 2019).

The result of mean time in solving each task with the four visualization techniques is shown in Fig. 6b. For T2 and T3, FuzzyRadar costs less time than the other three techniques. Thus, our multiple encodings can improve the efficiency of Radviz and PCP in analyzing the stability of data items. For T5, FuzzyRadar requires more time than Radviz and PCP. The volunteers claimed that they tried to count and compare the number of data points labeled in the colored area of histograms to obtain an accurate answer for dominant cluster identification, but they would not do this when using Radviz and PCP. This comment explains why FuzzyRadar costs more time and obtains better accuracy than Radviz and PCP for T5.

In terms of satisfaction, FuzzyRadar obtains high satisfaction scores for tasks T1, T2, T3, and T6 (Fig. 6c). For T4, FuzzyRadar outperforms Radviz but underperforms PCP and SPM, which reflects that the combination of Radviz and PCP can enhance the capability of Radviz in obtaining cluster information;

	FuzzyRadar	PCP	SPM	Radviz			FuzzyRadar	PCP	SPM	Radviz			FuzzyRadar	PCP	SPM	Radviz
T1	95.0%	93.3%	95.0%	96.7%		T1	11.7	16.17	16.57	12.5		T1	6.2	5.3	5.1	6.4
T2	93.3%	73.3%	76.7%	89.7%		T2	9.1	22.8	24.2	18.7		T2	6.5	4.8	4.6	6.3
T3	91.1%	77.8%	75.6%	88.8%		T3	21.2	39.3	40.1	24.6		T3	6.1	4.2	3.5	5.9
T4	89.1%	90.6%	79.7%	87.5%		T4	26.1	22.7	21.8	30.3		T4	5.7	6.2	6.1	5.2
T5	85.7%	78.5%	71.4%	82.2%		T5	48.5	38.1	49.5	31.6		T5	5.5	5.6	4.1	6.1
T6	93.3%	76.7%	68.9%	43.3%		T6	15.5	40.3	41.6	60.2		T6	6.5	5.2	5.1	4.1
T7	73.3%	43.3%	76.7%	56.7%		T7	102.5	110.3	61.2	80.3		T7	4.3	4.9	6.1	3.9
	(a)						**(b)**						**(c)**			

Fig. 6 Results of **a** mean accuracy, **b** mean time, and **c** mean satisfaction score in solving each task with the four visualization techniques. Colors indicate groups of no significant pairwise differences, with the winners shown in dark blue, losers in gray, and the ones in between in light blue. Taking T6 in **b** as an example, the four fields have various colors that reflect the significant differences; the dark blue FuzzyRadar is the winner, and the gray Radviz is the loser; whereas the fields of PCP and SPM are light blue, which indicates that they are in between and that no significant difference exists between them

however, the redesigned PCP causes a negative effect on recognizing detailed cluster information. For T7, FuzzyRadar obtains a low satisfaction score. The volunteers commented that they had difficulties in estimating the numbers of data items based on the widths of polygonal strips.

7 Discussion

In this section, we discuss the limitations of this work and suggest some interesting aspects for further work.

We do not include any data set with two clusters. Although two fuzzy clusters commonly exist in application scenarios, they are not typical multidimensional data and Radviz is unsuitable to visualize data with only two clusters.

The visual clutter caused by PCP polylines is significantly reduced using the proposed strip-edge-bundling method; however, a large number of data points may still lead to serious overlapping of data points within the FuzzyRadar circle. This problem can be solved by using the method in the literature (Artero and de Oliveira 2014; Novakova and Stepankova 2009) or combining the density map.

FuzzyRadar is unsuitable to present numerous clusters simultaneously. Numerous histograms outside the FuzzyRadar circle will occupy most of the screen space; thus, only little space for the polygonal strips remains. Moreover, a large number of clusters may result in the reordering problem of clusters. Generally, dimension reordering strategy has a significant influence on the visualization results of Radviz and PCP. Obtaining an ideal dimension order is difficult when dimensions are plenty. Available strategies include the methods mentioned in the literature (Leban et al. 2006; Albuquerque et al. 2010; Kuntal et al. 2014).

Although we provide triple visual encodings to help users identify dominant cluster, users still needed to estimate the number of data items belonging to each cluster on the basis of the maximum MD principle. Such estimation affects the accuracy of dominant cluster identification. This problem can be solved by displaying the number of selected data points or providing the size tips of clusters on polygonal strips.

When judging the correlations between two nonadjacent clusters, the polygonal strips between them may be interfered by histograms of other clusters. This interference may lead to inaccuracy in the judgment, which can be solved by hiding irrelevant histograms or allowing users to reorder the PCP axes.

This work currently concentrates on the seven analytical tasks of understanding fuzzy clusters. Many other analytical tasks about fuzzy clustering are worth studying. For example, we can introduce original data into the analysis of fuzzy clustering result. We can also develop some comparison functions to help users analyze multiple clustering results gained by different clustering parameters.

8 Conclusion

In this study, we design a new visualization called FuzzyRadar for understanding fuzzy clusters. In FuzzyRadar, a compact and compounded layout is proposed to combine Radviz and PCP. A strip-edge-bundling

method and a histogram embedding method are introduced to reduce visual cluster and facilitate the recognition of MD distribution. Additional visual encodings and a set of lightweight interactions are provided to help users perform a collaborative usage of Radviz and PCP. We use a case study to demonstrate the usability of FuzzyRadar and conduct a controlled quantitative evaluation to compare the performance of FuzzyRadar, Radviz, PCP, and SPM. The results show that FuzzyRadar generally outperforms Radviz, PCP, and SPM in completing all the seven examined analytical tasks of understanding fuzz clusters.

References

Abonyi J, Babuska R (2004) Fuzzsam-visualization of fuzzy clustering results by modified Sammon mapping. In: Proceedings of the IEEE international conference on fuzzy systems. IEEE, pp 365–370

Albuquerque G, Eisemann M, Lehmann DJ, Theisel H, Magnor M (2010) Improving the visual analysis of high-dimensional datasets using quality measures. In: Proceedings of the IEEE symposium on visual analytics science and technology, pp 19–26

Artero AO, de Oliveira MCF (2014) Viz3D: effective exploratory visualization of large multidimensional data sets. In: Proceedings of the 17th Brazilian symposium on computer graphics and image processing, pp 340–347

Asuncion A, Newman D (2007) UCI machine learning repository. http://archive.ics.uci.edu/ml/datasets.html. Accessed 15 Apr 2019

Berthold MR, Hall LO (2003) Visualizing fuzzy points in parallel coordinates. IEEE Trans Fuzzy Syst 11(3):369–374

Bi C, Fu B, Chen J, Zhao Y, Yang L, Duan Y, Shi Y (2019) Machine learning based fast multi-layer liquefaction disaster assessment. In: World Wide Web: internet and web information systems, pp 1–16. https://doi.org/10.1007/s11280-018-0632-8

Cao N, Lin YR, Gotz D (2015) UnTangle map: visual analysis of probabilistic multi-label data. IEEE Trans Vis Comput Graph 22(2):1149–1163

Chen H, Zhang S, Chen W et al (2015) uncertainty-aware multidimensional ensemble data visualization and exploration. IEEE Trans Vis Comput Graph 21(9):1072–1086

Chen W, Xia J, Wang X, Chen J, Wang Y, Chang L (2018a) RelationLines: visual reasoning of egocentric relations from heterogeneous urban data. ACM Trans Intell Syst Technol 10(1):1–22

Chen W, Huang Z, Wu F, Zhu M, Guan H, Maciejewski R (2018b) VAUD: a visual analysis approach for exploring spatio-temporal urban data. IEEE Trans Vis Comput Graph 24(9):2636–2648

De Oliveira JV, Pedrycz W (2007) Advances in fuzzy clustering and its applications. Wiley, London

Dimara E, Bezerianos A, Dragicevic P (2018) Conceptual and methodological issues in evaluating multidimensional visualizations for decision support. IEEE Trans Vis Comput Graph 24(1):749–759

Dunn JC (1973) A fuzzy relative of the ISODATA process and its use in detecting compact well separated cluster. J Cybern 3(3):32–57

Epp CD, Bull S (2015) Uncertainty representation in visualizations of learning analytics for learners: current approaches and opportunities. IEEE Trans Learn Technol 8(3):242–260

Etemadpour R, Motta R, Jg DSP (2015) Perception-based evaluation of projection methods for multidimensional data visualization. IEEE Trans Vis Comput Graph 21(1):81–94

Feil B, Balasko B, Abonyi J (2007) Visualization of fuzzy clusters by fuzzy sammon mapping projection: application to the analysis of phase space trajectories. Soft Comput 11(5):479–488

Ferreira N, Fisher D, Konig AC (2014) Sample-oriented task-driven visualizations: allowing users to make better, more confident decisions. In: Proceedings of the Sigchi conference. ACM, pp 571–580

Holten D, Van Wijk JJ (2010) Evaluation of cluster identification performance for different PCP variants. Comput Graph Forum 29(3):793–802

Hoppner F, Klawonn F (2006) Visualising clusters in high-dimensional data sets by intersecting spheres. In: Proceedings of the international symposium on evolving fuzzy systems. IEEE, pp 106–111

Inselberg A, Dimsdale B (1987) Parallel coordinates for visualizing multi-dimensional geometry. In: Proceedings of the international conference on computer graphics, pp 25–44

Klawonn F, Chekhtman V, Janz E (2003) Visual inspection of fuzzy clustering results. In: Proceedings of advances in soft computing. CRC, pp 65–76

Kuntal BK, Ghosh TS, Mande SS (2014) Igloo-Plot: a tool for visualization of multidimensional datasets. Genomics 103(1):11–20

Leban G, Zupan B, Vidmar G, Bratko I (2006) Vizrank: data visualization guided by machine learning. Data Min Knowl Discov 13(2):119–136

Lin YR, Cao N, Gotz D, Lu L (2015) Untangle: visual mining for data with uncertain multi-labels via triangle map. In: Proceedings of the IEEE international conference on data mining. IEEE, pp 340–349

Liu M, Shi J, Li Z, Li C, Zhu J, Liu S (2017) Towards better analysis of deep convolutional neural networks. IEEE Trans Visual Comput Graph 23(1):91–100

Long TV (2014) iSPLOM: interactive with scatterplot matrix for exploring multidimensional data. In: Proceedings of the international conference on knowledge and systems engineering, pp 175–186

Ma Y, Tung AKH, Wang W, Gao X, Pan Z, Chen W (2019) ScatterNet: a deep subjective similarity model for visual analysis of scatterplots. IEEE Trans Vis Comput Graph. https://doi.org/10.1109/tvcg.2018.2875702

Manuel R, Laura R, Francisco D, Alberto S (2016) A comparative study between Radviz and star coordinates. IEEE Trans Vis Comput Graph 22(1):619–628

Marghescu D (2007) User evaluation of multidimensional data visualization techniques for financial benchmarking. In: Proceedings of the European conference on information management and evaluation. AMCIS, p 509

Novakova L, Stepankova O (2009) Radviz and identification of clusters in multidimensional data. In: Proceeding of the 17th international conference on information visualization, pp 104–109

Palmas G, Bachynskyi M, Oulasvirta A (2014) An edge-bundling layout for interactive parallel coordinates. In: Proceedings of IEEE Pacific visualization symposium. IEEE, pp 57–64

Peng W, Ward MO, Rundensteiner EA (2004) Clutter reduction in multidimensional data visualization using dimension reordering. In: Proceedings of the IEEE symposium on information visualization. IEEE, pp 89–96

Rueda L, Zhang Y (2006) Geometric visualization of clusters obtained from fuzzy clustering algorithms. Pattern Recogn 39(8):1415–1429

Ruspini EH (1969) A new approach to clustering. Inf Control 15(15):22–32

Rzeźniczak T (2013) Evaluation of multidimensional visualization techniques for medical patterns representation. J Theor Appl Comput Sci 7(4):75–85

Sanyal J, Zhang S, Bhattacharya G (2009) A user study to compare four uncertainty visualization methods for 1D and 2D datasets. IEEE Trans Vis Comput Graph 15(6):1209–1218

Sedlmair M, Munzner T, Tory M (2013) Empirical guidance on scatterplot and dimension reduction technique choices. IEEE Trans Vis Comput Graph 19(12):2634–2643

Sharko J, Grinstein G (2009) Visualizing fuzzy clusters using RadViz. In: Proceedings of the 13th international conference on information visualisation. IEEE, pp 307–316

Shi R, Yang M, Zhao Y, Zhou F, Huang W, Zhang S (2016) A matrix-based visualization system for network traffic forensics. IEEE Syst J 10(4):1350–1360

Shi Y, Bryan C, Bhamidipati S, Zhao Y, Zhang Y, Ma K-L (2018) MeetingVis: visual narratives to assist in recalling meeting context and content. IEEE Trans Vis Comput Graph 24(6):1918–1929

Wu Y, Yuan G-X, Ma K-L (2012) Visualizing flow of uncertainty through analytical processes. IEEE Trans Vis Comput Graph 18(12):2526–2535

Wu Y, Cao N, Gotz D, Tan Y-P, Keim DA (2016) A survey on visual analytics of social media data. IEEE Trans Multimed 18(11):2135–2148

Xia J, Ye F, Chen W, Wang Y, Chen W, Ma Y, Tung AKH (2018a) LDSScanner: exploratory analysis of low-dimensional structures in high-dimensional datasets. IEEE Trans Vis Comput Graph 24(1):236–245

Xia J, Gao L, Kong K, Zhao Y, Chen Y, Kui X, Liang Y (2018b) Exploring linear projections for revealing clusters, outliers, and trends in subsets of multi-dimensional datasets. J Vis Lang Comput 48:52–60

Xie C, Chen W, Huang X, Hu Y, Barlowe S, Yang J (2014) VAET: a visual analytics approach for E-Transactions time-series. IEEE Trans Vis Comput Graph 20(12):1743–1752

Yang J, Peng W, Ward M, Rundensteiner E (2003) Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In: Proceedings of IEEE symposium on information visualization. IEEE, pp 105–112

Zadeh LA (1965) Fuzzy sets. Inf Control 8(3):338–353

Zhao Y, She Y, Chen W, Lu Y, Xia J, Chen W, Liu J, Zhou F (2018) EOD edge sampling for visualizing dynamic network via massive sequence view. IEEE Access 6(1):53006–53018

Zhao Y, Luo F, Chen M, Wang Y, Xia J, Zhou F, Wang Y, Chen Y, Chen W (2019) evaluating multi-dimensional visualizations for understanding fuzzy clusters. IEEE Trans Vis Comput Graph 25(1):12–21

Zhou F, Li J, Huang W (2016) Dimension reconstruction for visual exploration of subspace clusters in high-dimensional data. In: Proceedings of the IEEE Pacific visualization symposium. IEEE, pp 128–135

Zhou Z, Ye Z, Liu Y, Liu F, Tao Y, Su W (2017a) Visual analytics for spatial clusters of air-quality data. IEEE Comput Graph Appl 37(5):98–105

Zhou F, Chen M, Wang Z (2017b) A Radviz-based visualization for understanding fuzzy clustering results. In: Proceedings of the international symposium on visual information communication and interaction, pp 9–15

Zhou F, Lin X, Liu C, Zhao Y, Xu P, Ren L, Xue T, Ren L (2019a) A survey of visualization for smart manufacturing. J Vis 22(1):1–19

Zhou Z, Meng L, Tang C, Zhao Y, Guo Z, Miaoxin H, Chen W (2019b) Visual abstraction of the large scale geospatial origin-destination movement Data. IEEE Trans Vis Comput Graph 25(1):43–53. https://doi.org/10.1109/TVCG.2018.2864503

Zuk T, Carpendale S (2006) Theoretical analysis of uncertainty visualizations. In: Proceedings of the international society for optical engineering, vol 6060, pp 606007–606014