Visual Approach to Boundary Detection of Clusters Projected in 2D Space

105

Lenon Fachiano Silva and Danilo Medeiros Eler

Abstract

Data mining tasks are commonly employed to aid users in both dataset organization and classification. Clustering techniques are important tools among all data mining techniques because no class information is previously necessary – unlabeled datasets can be clustered only based on their attributes or distance matrices. In the last years, visualization techniques have been employed to show graphical representations from datasets. One class of techniques known as multidimensional projection can be employed to project datasets from a high dimensional space to a lower dimensional space (e.g., 2*D* space). As clustering techniques, multidimensional projection techniques present the datasets relationships based on distance, by grouping or separating cluster of instances in projected space. Usually, it is difficult to detect the boundary among distinct clusters presented in 2*D* space, once they are projected near or overlapped. Therefore, this work proposes a new visual approach for boundary detection of clusters projected in 2*D* space. For that, the attributes behavior are mapped to graphical representations based on lines or colors. Thus, images are computed for each instance and the graphical representation is used to discriminate the boundary of distinct clusters. In the experiments, the color mapping presented the best results because it is supported by the user's pre-attentive perception for boundary detection at a glance.

Keywords

Text mining • Document pre-processing • Visualization • Document similarity • Multidimensional projection

105.1 Introduction

Clustering algorithms can be used to divide a dataset in distinct groups (or clusters), enabling users to focus in specific groups or to organize a dataset to facilitate its exploration and understanding. A main issue in most clustering algorithms is how to specify the number of groups to cluster a dataset. Alternatively, some visualization techniques can be employed to present the dataset instances similarities in 2*D* space, thus, the groups are naturally created according to the feature space that describes the dataset. Usually, multidimensional projection techniques [13] are employed in this visualization process by reducing the dataset dimensionality to two dimensions.

Even though multidimensional projection techniques can represent structures from original multidimensional space in the projected space, groups can be overlapped or share a common boundary. The first case is a consequence of a dimensionality reduction and projection techniques. In the second case, when groups of distinct classes of instances

L.F. Silva (✉) • D.M. Eler
Faculdade de Ciências e Tecnologia, Departamento de Matemática e Computação, UNESP – Universidade Estadual Paulista, Presidente Prudente, SP, Brazil
e-mail: lenon_fachiano@hotmail.com; daniloeler@fct.unesp.br

S. Latifi (ed.), *Information Technology – New Generations*, Advances in Intelligent Systems and Computing 558, DOI 10.1007/978-3-319-54978-1_105

share a common border, the main issue is to identify the boundary of different clusters. In the literature, distinct approaches can be employed to identify boundary among clusters. Usually, they are based on clustering algorithms employed in 2*D* space [11]; edge connecting based on similarities or triangulations performed in 2*D* space [10]; approaches for understanding projection and attributes dimensions based on most representative attributes [3]; hybrid visualizations to explain attribute behaviours [1, 5]; and other ways of cluster identification and space division [8, 9]. Even though they can highlight a border among clusters, the presented boundary do not reveal the real boundary from the wholly feature space, but a frontier imposed by 2*D* space – in the case of clustering algorithms employed in 2*D* space or space division – or based on some attributes – in the case of highlighting of the most representative attributes.

In this paper, we propose a new approach for boundary detection of clusters projected in 2*D* space. For that, we use multidimensional projection for cluster formation and instances relationship mapping in 2*D* space (i.e., projected space). We transform this similarity based exploration in a hybrid exploration by adding other class of visualization techniques based on attributes analysis. Thus, after mapping the instances similarities for 2*D* space, our hybrid approach generates an image for each instance to highlight the attributes behaviour, enabling the user to differentiate instances from distinct classes that share a common boundary. In this paper we propose two kinds of attributes mapping: one based on lines, in which polylines are computed from attribute values like in Parallel Coordinates technique [7]; and other based on color, in which each attribute value is mapped for a color.

The main contribution of this paper is to aid the dataset exploration based on a new hybrid exploration approach capable of highlight boundary of distinct clusters projected in 2*D* space. Additionally, this approach uses all dataset attributes to generate the visual representations and can also be used to understand the cluster formation based on the attributes analysis. Thus, the feature space is understood based on the similarities mapping from projection techniques and attributes mapping based on our approach.

This paper is organized as follows. Section 105.2 presents the theoretical foundation of visualization techniques employed in this work: multidimensional projection, parallel coordinates and color mapping. Section 105.3 presents the boundary detection approach proposed in this work. Section 105.4 presents the performed experiments with the proposed approach employed to boundary detection and feature space comprehension. Section 105.5 concludes the paper, summarizing the main achievements and projecting further works.

105.2 Background

Visualization is a research area whose main goal is to enable exploration, understanding and analysis of datasets through interactive visual explorations [2]. Visualization techniques generate graphical representations to facilitate the user comprehension and perception during a dataset exploration [4, 14]. This work is based on three visualization techniques: multidimensional projection, parallel coordinates and color mapping.

The multidimensional projection technique aims to map the instances similarities from the original space in a lower dimensional space. The dataset instances are mapped according to the similarities relationships from the original multidimensional space [11–13]. Thus, groups and neighbourhood from the original space are kept in the lower dimensional space – in this work, they are projected in 2*D* space. Scatter plots are used as graphical representations, but, instead of using two attributes from the dataset, the X and Y positions are computed based on all attributes that are projected to two dimensions.

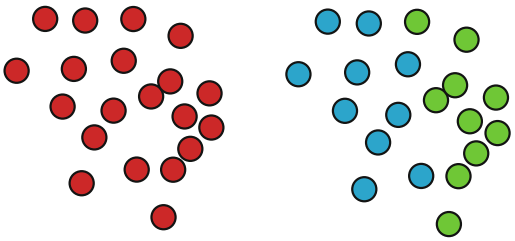
The line mapping presented in this work is based on Parallel Coordinates technique [7] which main objective is to show the attributes behaviour and relationship. For that, Parallel Coordinates uses A_k parallel axes representing K attributes from a dataset – each axes assigned to an attribute. The instances are mapped as polylines that intersect each axes at the value of the attribute K_i for the axes A_i. Thus, the user can compare attributes behaviour by looking at the line shapes. In this work, instead of drawing all lines, we only draw one line for each instance.

The last technique employed in this work is a classical color mapping that aims to map data values to color according to a predefined color scale. Thus, each attribute value is mapped to a color, in this case, the user looks to the color information to understand the attribute behaviour.

Next section presents the proposed approach for boundary detection.

105.3 Boundary Detection Approach

This section describes the proposed boundary detection approach based on line and color mapping. This approach can be employed to detect clusters frontier in the projected space. For instance, in Fig. 105.1a is shown a group of instances in 2*D* space, but the label information is not available. So, detecting the boundary between distinct groups is to laborious, needing other techniques to improve this detection task. As shown in Fig. 105.1b, only with label information it is possible to perfectly detect the boundary between these two clusters.



(a) No label information

(a) Colored with label information

Fig. 105.1 Example of clusters in projected space: (**a**) no label information provided and (**b**) instances are colored based on label information

Our proposed approach aid in boundary detection based on color and line mappings from attributes information of each dataset instance. An example of mapping from three distinct feature vectors is shown in Fig. 105.2. The color mapping is based on a previous defined color scale by means which each attribute value is mapped to a color. Thus, an image is generated for each instance and divided in columns according to the number of attributes. Each column is colored with the color mapping for the respective attribute value. Similarly, for the line mapping, an image is generated with a polyline to represent the attributes behaviour for the respective instance. The polyline is generated based on Parallel Coordinates [7] technique, in which the space is divided in parallel axes – one axis for each attribute – and a polyline intersects these axes in the point correspondent to the attribute value.

Next section show examples of this approach in boundary detection of a dataset.

105.4 Experiments

This section presents experiments to show how the proposed approach can be employed in boundary detection of datasets projected in 2*D* space. The first experiment used the motivating example presented in Fig. 105.1. For that, the attributes information was used to generate images for each instance based on color and line mappings. Figure 105.3 presents the results of these mappings. The boundary detection based on lines can be performed by looking at the changing in specific attributes behaviour, for instance, we can note in each image that the values of some attributes of the right cluster is lower than the left cluster. On the other hand, the boundary detection based on color is more precise and evident. The left cluster present blue color mapping for some attributes and right cluster present red color mapping for the same attributes.

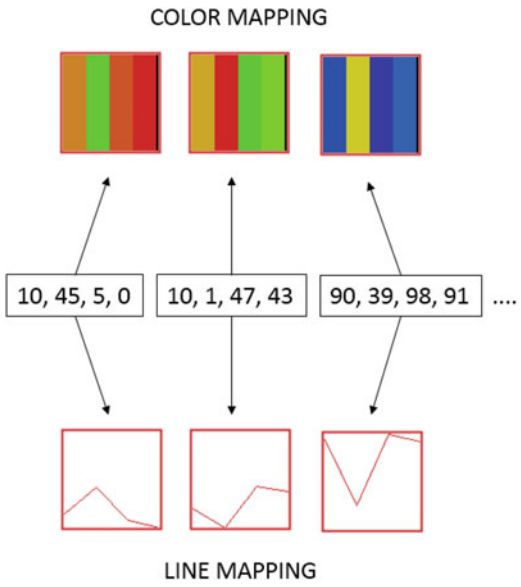


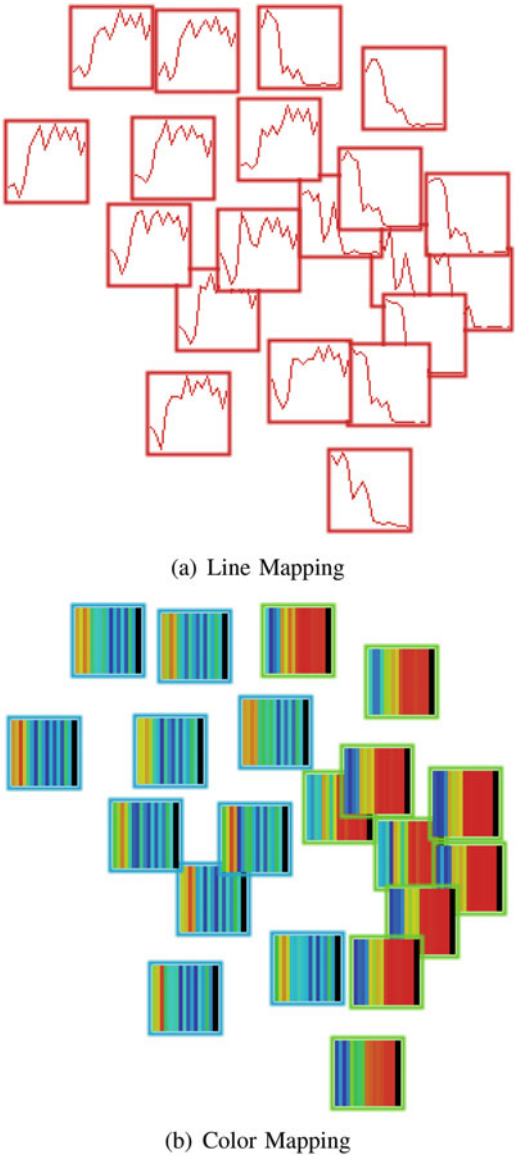
Fig. 105.2 Boundary detection approach

Other experiment was performed with a dataset composed by three classes – the well known Iris dataset.[1] Figure 105.4a shows the 2*D* representation of this dataset, in which is not possible to detect the clusters boundary. After applying our approach, color and line mapping were computed. Firstly, by analysing the clusters boundary based on line mapping, as show in Fig. 105.4b, we can detect a left cluster whose boundary was indicated by a blue line. However, the detection of other two clusters transition at right is to laborious when using the line mapping. On the other hand, by using the color mapping, as shown in Fig. 105.4c, the left cluster is evident and the blue line indicates its boundary. The color mapping facilitated the detection of other clusters boundary as indicated by the red line. To validate our approach, in Fig. 105.4d is shown the color information based on instances label (class) as a ground truth.

These experiments show that our approach can be employed to aid in boundary detection of clusters projected in 2*D* space. Additionally, the perception based on color information enables users to rapidly and accurately detect a boundary among groups of elements.

The proposed approach can also be employed to show bad feature spaces, that is, feature space that cannot discriminate instances from distinct classes. To show this application, we consider an image dataset with MRI images from six distinct classes. As shown in Fig. 105.5a, we can note some groups based on the projected clusters and detect the boundary of some clusters based on color mapping.

[1] Iris dataset is available at http://archive.ics.uci.edu/ml/datasets/Iris



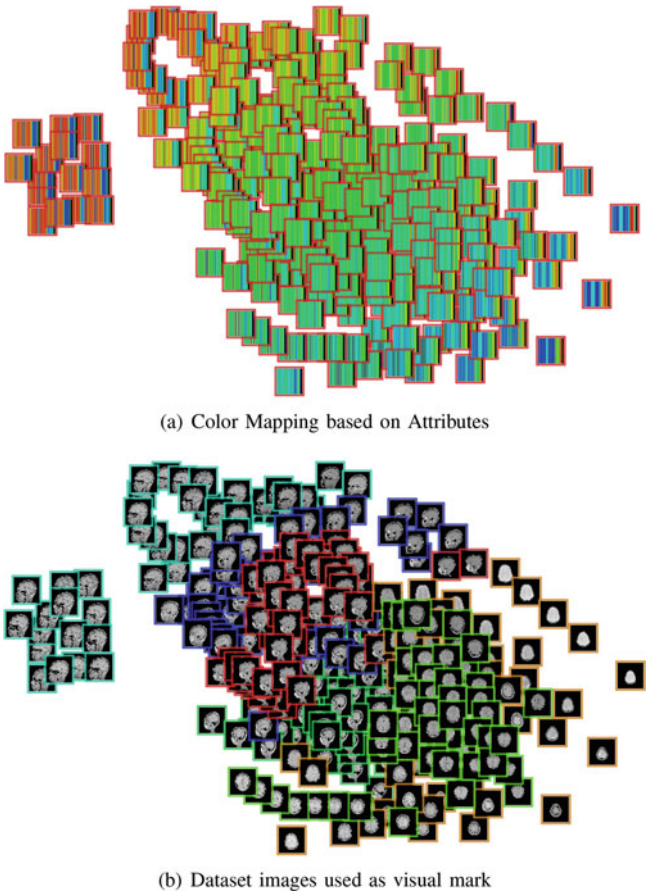
(a) Line Mapping

(b) Color Mapping

Fig. 105.3 Boundary detection using our approach by mapping the attributes values for line (**a**) and color (**b**)

However, the boundary detection is too laborious at the centre of the big cluster, where the green color is similar to all attributes of several images. Using the dataset images as visual marks [6] we can perceive that the features (attributes) employed to this dataset is not able to discriminate instances of different classes. There are four classes of head that are too similar. Therefore, these images are considered as a big cluster as shown based on the color mapping. Thus, it is necessary to use other kind of features to separate this classes of images or considering it all as misclassified from the expert.

Next section presents some conclusions and further works.



(a) Color Mapping based on Attributes

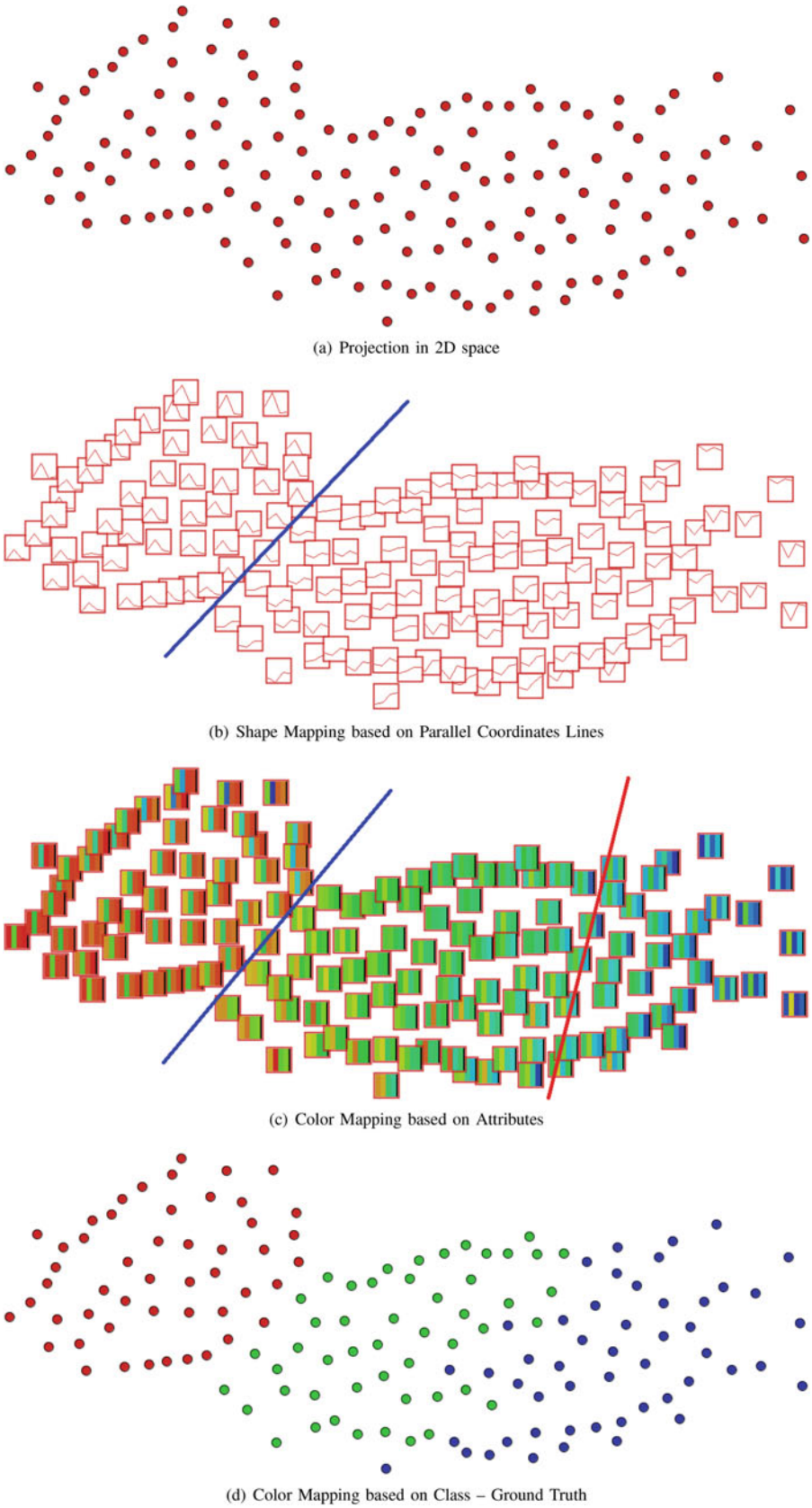
(b) Dataset images used as visual mark

Fig. 105.4 Boundary detection of a dataset projected in 2D space (**a**). Our approach was employed to show two mappings based on shape (**b**) and color (**c**). In (**d**) is shown the real class information from the dataset

105.5 Conclusions and Future Works

Boundary detection among distinct cluster is an important step in dataset organizations. In this work, we presented a visual approach based on multidimensional projection technique to reduce the dataset dimensionality and project it in $2D$ space. Thus, clusters are presented based on instances similarities from the original multidimensional space. To identify the cluster boundaries, we used line and color mappings to show the attributes relationship and highlight the cluster boundaries.

In the experiments, we used the line mapping based on Parallel Coordinates to show the attributes behaviour, by means which the user could identify the transition among clusters by looking at the shape of each line. However, the shape comparison can be too laborious. Therefore, color mapping was also employed as visual characteristic and the boundaries could be rapidly perceived.



(a) Projection in 2D space

(b) Shape Mapping based on Parallel Coordinates Lines

(c) Color Mapping based on Attributes

(d) Color Mapping based on Class – Ground Truth

Fig. 105.5 Boundary detection of a dataset projected in 2D space based on color mapping (**a**). In (**b**) the corresponding dataset images were employed as visual mark

Color mapping is more effective in boundary detection tasks because the hue is a pre-attentive visual property. As shown in the results, boundaries could be detected at a glance with color mapping. While the line (shape) requires an attentive user perception, that is, the user needs to focus and search for the boundary detection by comparing line shapes. In further works, other pre-attentive visual properties could be employed as well as combination of shape and color in the same image.

Besides the boundary detection, the last experiment presented in previous section shows that our approach can also be employed to explain the cluster cohesion and separation. Thus, experts can analyse the attributes behaviour in cluster formation and decide the quality of the feature spaces as well as identify some wrong in instances labelling task.

The main limitation of this approach is the size of generated images, because when the dataset have more attributes than the image size in pixels, we cannot generate an image with all attributes – for instance, in text mining, is common a feature space with more than 500 attributes. Therefore, in further works, approaches to feature selection can be employed to show attributes behaviour of datasets with several attributes.

Acknowledgements The authors acknowledge the financial support of the Brazilian financial agency São Paulo Research Foundation (FAPESP) – grant #2013/03452-0, National Counsel of Technological and Scientific Development (CNPq), and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

References

1. Bodo, L., de Oliveira, H. C., Breve, F. A., & Eler, D. M. (2016). Performance indicators analysis in software processes using semi-supervised learning with information visualization. In *13th International Conference on Information Technology: New Generations (ITNG 2016)* (Advances in intelligent systems and computing, pp. 555–568). Las Vegas, NV: Springer.
2. Card, S., Mackinlay, J., & Shneiderman, B. (1999). *Readings in information visualization: Using vision to think* (Interactive technologies Series). San Francisco, CA: Morgan Kaufmann Publishers.
3. da Silva, R. O., Rauber, P. E., Martins, R. M., Minghim, R., & Telea, A. C. (2015). Attribute-based visual explanation of multidimensional projections. In *EuroVis workshop on visual analytics (EuroVA)*, Cagliari. The Eurographics Association.
4. de Oliveira, M., & Levkowitz, H. (2003). From visual data exploration to visual data mining: a survey. *IEEE Transactions on Visualization and Computer Graphics, 9*(3), 378–394.
5. de Oliveira, R. A. P., Silva, L. F., & Eler, D. M. (2015). Hybrid visualization: A new approach to display instances relationship and attributes behaviour in a single view. In *19th International Conference on Information Visualisation (iV)*, Barcelona (pp. 277–282).
6. Eler, D. M., Nakazaki, M. Y., Paulovich, F. V., Santos, D. P., Andery, G. F., Oliveira, M. C. F., Batista Neto, J., & Minghim, R. (2009). Visual analysis of image collections. *The Visual Computer, 25*(10), 923–937.
7. Inselberg, A. (1985). The plane with parallel coordinates. *The Visual Computer, 1*(2), pp. 69–91.
8. Kandogan, E. (2012). Just-in-time annotation of clusters, outliers, and trends in point-based data visualizations. In *Proceedings of the 2012 I.E. Conference on Visual Analytics Science and Technology VAST'12* (pp. 73–82). Washington, DC: IEEE Computer Society.
9. Nocaj, A., & Brandes, U. (2012). Organizing search results with a reference map. *IEEE Transactions on Visualization and Computer Graphics, 18*(12), 2546–2555.
10. Paulovich, F. V., Nonato, L. G., & Minghim, R. (2006). Visual mapping of text collections through a fast high precision projection technique. In *Proceedings of the Conference on Information Visualization IV'06* (pp. 282–290). Washington, DC: IEEE Computer Society.
11. Paulovich, F. V., Oliveira, M. C. F., & Minghim, R. (2007). The projection explorer: A flexible tool for projection-based multidimensional visualization. In *XX Brazilian Symposium on Computer Graphics and Image Processing, SIBGRAPI 2007* (pp. 27–36). IEEE, Belo Horizonte.
12. Paulovich, F. V., Nonato, L. G., Minghim, R., & Levkowitz, H. (2008). Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. *IEEE Transactions on Visualization and Computer Graphics, 14*(3), 564–575.
13. Tejada, E., Minghim, R., & Nonato, L. G. (2003). On improved projection techniques to support visual exploration of multidimensional data sets. *Information Visualization, 2*(4), 218–231.
14. Ware, C., (2012). *Information visualization: Perception for design* (Interactive technologies). Amsterdam: Elsevier Science.