Analyzing the role of dimension arrangement for data visualization in Radviz

Analyzing The Role Of Dimension Arrangement For Data Visualization in Radviz

Luigi Di Caro[1], Vanessa Frias-Martinez[2], and Enrique Frias-Martinez[2]

[1] Department of Computer Science, Universita' di Torino, Torino, Italy
dicaro@di.unito.it
[2] Data Mining and User Modeling Group, Telefonica Research, Madrid, Spain
{vanessa,efm}@tid.es

Abstract. The Radial Coordinate Visualization (Radviz) technique has been widely used to effectively evaluate the existence of patterns in highly dimensional data sets. A crucial aspect of this technique lies in the arrangement of the dimensions, which determines the quality of the posterior visualization. Dimension arrangement (DA) has been shown to be an NP-problem and different heuristics have been proposed to solve it using optimization techniques. However, very little work has focused on understanding the relation between the arrangement of the dimensions and the quality of the visualization. In this paper we first present two variations of the DA problem: (1) a Radviz independent approach and (2) a Radviz dependent approach. We then describe the use of the Davies-Bouldin index to automatically evaluate the quality of a visualization *i.e.,* its visual usefulness. Our empirical evaluation is extensive and uses both real and synthetic data sets in order to evaluate our proposed methods and to fully understand the impact that parameters such as number of samples, dimensions, or cluster separability have in the relation between the optimization algorithm and the visualization tool.

1 Introduction

Visualization tools focus on graphically representing high dimensional and multivariate data with enough clarity to allow for data exploration. Low dimensional data sets have traditionally been represented using either simple line graphs or scatter plots. Nevertheless, in the case of high dimensional data sets, special techniques for data visualization such as Parallel Coordinates [6], Star Glyphs [7], Circle Segments [2] or Radviz [12] are used. One of the key problems of these techniques is the dimension arrangement problem (DA), which evaluates from an algorithmic perspective which arrangement of the dimensions facilitates more the comprehension of the data. Ankerst *et. al* [1] formalized the DA problem and proved that it is NP-complete similarly to the traveling salesman problem. In this paper we present two reformalization of it designed to explore a search space whose non-convexity makes it more probable to find the desired global maxima (minima). The evaluation of the effectiveness of the arrangement in terms of visual information is typically carried out by means of human intervention. Most

of the papers focusing on visualization techniques have generally assumed that the better the solution for the dimension arrangement optimization problem, the better the visual usefulness of the projected data. In this paper, we present an initial approach to formally determine such relation, making use of the Davies-Bouldin index for cluster analysis in order to compute the visual quality of the information being plotted by Radviz by an extensive empirical evaluation on synthetic and real datasets.

2 Related Work

There is a wide variety of visualization techniques for multidimensional data that present a circular arrangements of the dimensions, like Star Coordinates [7], Circle Segments [2] and Circle Graphs [11]. We focus our analysis on Radviz [4] which we further explain in Section 3. The problem of dimension arrangement is common for all circular and non-circular visualization techniques and was formalized by Ankerst *et al.* as an optimization problem where the similarity between dimensions located next to each other had to be maximized. to be NP-complete. So far, very little work has been done to automatically understand (without human intervention) the quality of the visualization for the projected data. Ankerst *et al.* evaluate the *goodness* of their dimension arrangement algorithms by simply stating that the results show clearly superiority. Yang *et al.* [13] proposed an interactive hierarchical ordering of the dimensions based on their similarities, thus improving the manageability of high-dimensional datasets and reducing the complexity of the ordering. Weng *et al.* [10] formalize the concept of clutter in various visualization techniques and present it as a dimension arrangement optimization problem whose solutions will improve the detection of structure in the projected data. Yang *et. al* [14] present a visualization technique where the user can interactively navigate through the dimensions and visually determine the quality of the re-arrangement. VizRank [9] is one of the few works that attempts to automate the visual quality evaluation process, by assessing data projections and ranking them according to their ability to visually discriminate between classes. The quality of the class separation is estimated by computing the predictive accuracy of the k-nearest neighbour classifier. Our evaluation scheme is faster and simpler than the VizRank approach and does not suffer from the typical k-NN problems such as the computation of an adequate value for k or the computational complexity ($O(n^2)$).

3 Radviz's algorithm

RadViz (Radial Coordinate visualization) [4][5] is a visualization technique based on Hooke's law that maps a set of n-dimensional points into a plane: each point is held in place with springs that are attached at the other end to the feature anchors. The stiffness of each spring is proportional to the value of the corresponding feature and the point ends up at the position where the spring forces are in equilibrium. Prior to visualization, feature values are scaled to lie between

0 and 1. Radviz offers a unique method which can help to identify relations among data. Its main advantage is that it needs no feature projections and provides a global view on the multidimensional, multivariate data. The condition of equilibrium for a single object $\boldsymbol{u}$ is given by $\sum_{i=0}^{n-1}(\boldsymbol{A}_i - \boldsymbol{u}) * y_i = 0$.

Radviz faces several open problems: *overlapping* (different objects can be placed in the same 2D point), *visual clutter* (different instances could be placed close to each other) and *NP-completeness* (the final effectiveness of the approach depends on the dimension arrangement). Despite that, no study has shown yet whether there exists a relation between the solution provided by the optimization algorithm and the improvement in the visual usefulness of the projection.

4 Dimension Arrangement Formalizations

Although the DA problem has already been formalized in a generic context by Ankerst *et al.* [1], here we present new formalizations within the context of Radviz with the goal of providing a better exploration of the search spaces.

4.1 Independent DA

Let us assume that we have a dataset with points m that represent information represented with d dimensions. We define the similarity matrix as a symmetric matrix of dimensions $d \times d$, where each element $S_{i,j}$ represents the similarity between dimensions i and j. Each dimension i is represented as a distribution of m elements, where each element is taken from the $i-th$ dimension of each point in the dataset. In the experimental section we will describe the various metrics we have used to compute such similarity metric. Additionally, we define the neighborhood matrix N of dimensions $d \times d$which describes the circular distance between any two dimensions located in the circle. In particular, we calculate each $N_{i,j}$ as $1 - \frac{cdist(i,j)}{(d/2)}$, where d is the total number of dimensions and $cdist(i,j)$ represents the circular distance between dimensions i and j located on the circle. This distance is calculated as the number of dimensions on the circle between i and j through the shortest circular path. The larger the value of $N_{i,j}$, the closer the dimensions i and j are on the circle.

Thus, we can then formalize the dimension arrangement problem for a pair of similarity matrix S and neighborhood matrix N as a maximization problem where $\sum_{i=0}^{d-1}\sum_{j=0}^{d-1} N_{i,j} * S_{i,j}$ achieves its maximum value (*i.e.,* the more similar two dimensions are, the closer they should be located in the arrangement.

4.2 Radviz-dependent DA

Our second DA formalization focuses on using Radviz to evaluate the quality of the arrangement. Again, we start with the similarity matrix S of dimensions $d \times d$. For each possible dimension arrangement, this matrix represents a measure of the similarities across dimensions. For each specific matrix S, we project each row in S onto the circle using Radviz. The idea is that the projected dimension should

be as close to its dimension on the circle as possible. If that does not happen, it may be either that the dimensions are highly correlated or that the dimension arrangement is not good. Thus, for each dimension arrangement, each dimension i in the graph will have two representations: its coordinates on the circle, and its projected coordinates inside the circle, where the arranged dimensions are located according to the angular positions and the projected dimensions are calculated with respect to the Radviz formula.

Thus, the dimension arrangement problem can be defined as an optimization problem where for a given similarity matrix S, the optimal dimension arrangement is given by minimizing the sum of Euclidean distances between the arranged and the projected dimensions within the graph. This formalization follows the fact that the shorter the distance between an arranged dimension and its projection, the better the quality of the arrangement.

5 Experimental Setting

We want to focus our analysis on the relationship between the multiple DAs, the optimization functions, and the quality of visualization. For that purpose, in our analysis we make use of datasets with a limited number of dimensions that will allow us to fully explore, through a brute-force analysis, all the range of possible solutions. Our aim is twofold:

- To understand whether the formalization of the optimization problem as well as the metrics to measure similarity play a role in the way the search space (of the dimension arrangements) is explored.
- To carry an extensive experimental evaluation with both real and synthetic datasets to determine the relationship between the dimension arrangements and the quality of their associated data projections, studying the impact of various parameters like number of instance, dimensions, classes, and overlapping of the classes.

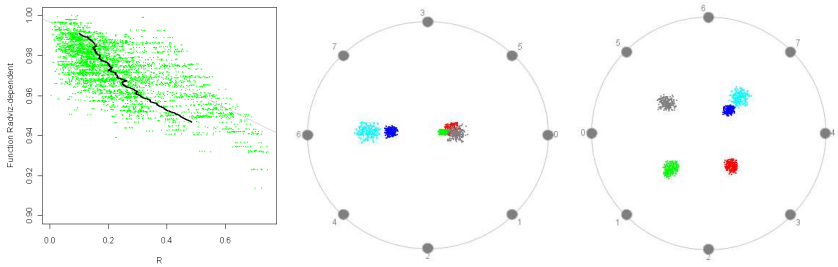
Regarding the synthetic data generation, we define four parameters for our algorithm: the number of classes *nc* (values from 2 to 100); the number of dimensions *nd*[1]; the number of instances *ni* (values from 100 to 10000) and the percentage *p_overlap* (up to 40%)[2] of instances that are randomly moved from one class to another. For each possible combination of *nc*, *nd* and *ni*, we create random instances within each class such that the clusters representing the classes are initially separated by equal distances. Finally, we modify the membership of a percentage *p_overlap* of the instances such that the boundaries between classes become blurry and classes start to overlap. We then used several real datasets from the UCI Machine Learning Repository[3].

The DA formalizations we have proposed are based on similarity measurements between dimensions. Although there exist many metrics to measure the

[1] We imposed a max # of dimensions (*8*) to be able to fully explore all possible DAs.
[2] Larger values did not add any extra overlap and were not considered.
[3] Datasets available at http://archive.ics.uci.edu/ml/datasets.html

similarity, we make use of the the Kullback-Leibler divergence [15] and the Cosine Similarity. The Kullback-Leibler (KL) divergence measures the difference between two probability distributions P and Q with cardinality d [4]: $\sum_{i=1}^{d} P_i * log_2(\frac{P_i}{Q_i})$. The inverse of it represents the similarity between them. On the other hand, Cosine similarity is calculated as the dot product between the distributions P and Q divided by the product of their norms. In order to study the relationship between a dimension arrangement and the visual usefulness of its projected data in Radviz, we first need to determine how visual usefulness is measured. The quality of the projected data onto the circle is related to the quality of the clusters obtained *i.e.,* the better the separation across clusters and instances, the more information the visual representation will convey to the data analyst. Thus, we measure visual usefulness of a data projection (and its corresponding dimension arrangement) using the Davies-Bouldin index (DB) [3]. DB is known to be one of the best indices to measure both the inter- and intra-cluster separation [8]. The DB index is computed as $\frac{1}{n} * \sum_{i=1}^{n} \overset{i \neq j}{max} \left\{ \frac{S_n(Q_i)+S_n(Q_j)}{S_n(Q_i,Q_j)} \right\}$, where n is the number of clusters, $S_n(Q_i)$ is the average Euclidean distance from each instance in cluster Q_i to its cluster centroid, and $S(Q_i, Q_j)$ is the Euclidean distance between cluster centers (Q_j can be any one of the clusters, a part from Q_i). Hence, the smaller the ratio, the more compact and separated the clusters are. Consequently, we seek dimension arrangements whose corresponding data projections have small DB indices associated. However, it may be the case that an initial dataset of instances with d dimensions shows a very high DB index in the d dimensional space, and thus it becomes very hard for its projected dataset to offer a good visualization. Thus, instead of measuring the DB of a projection, we measure the ratio R between the DB in the original data and the DB in the 2-dimensional mapping. Higher values of R correspond to higher visualization quality of the projected data. The first objective of the experimental evaluation is to be able to determine the relationship between the dimension arrangement and the quality of the associated visualization for each combination of the following parameters: (i) a specific dataset, either real or synthetic, (ii) a specific formalization of the DA problem, and (iii) a specific metric. Figure 1(a) shows an example result with the function *Radviz-dependent*. The number of points represents the number of dimension arrangements, while the black line represents the average values of a sliding window that captures the trend of density. We can observe that low values of the optimization function correspond to high values of R (minimization problem). The relation will be inverse when considering the function *independent* (maximization problem). Figures 1(b) and 1(c) show the Radviz projections associated to the worst value of R and the best value of R respectively. In Figure 2(a) we can observe that as the number of samples in the initial dataset increases, the best visual quality values R for the projected data decreases logarithmically. Figure 2(b) shows the visual information R versus the value of the optimization function for all possible DAs of datasets with 4, 6, and 8 dimensions. We can infer a general trend whereby as the number

[4] We used a symmetric version of the original Kullback-Leibler divergence.

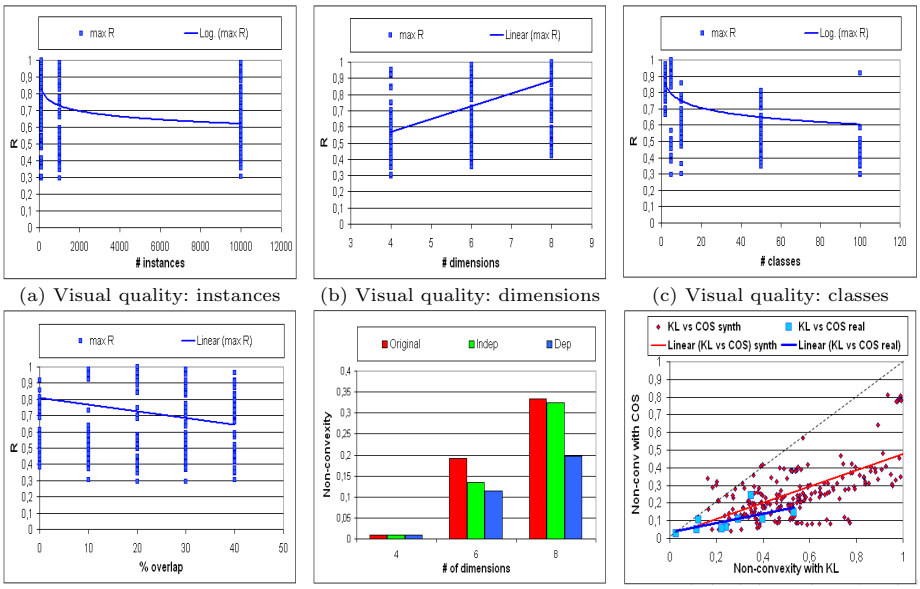


(a) Trend (b) Worst R (c) Best R

Fig. 1. (a) shows the correlation between the optimization function *Radviz-dependent* and the visual usefulness R of the DAs (the green points), using a synthetic dataset with 5 classes, 1000 instances, 8 dimensions and 10% of overlap. (b) and (c) show the projections with the best and the worst value of R.

of dimensions increases, the visual usefulness also improves following a linear curve. This result implies that as the number of dimensions grow, the Radviz technique manages to better maintain the initial distribution of the dataset *i.e,* the more dimensions, the better the samples can be characterized and the better Radviz will perform. Furthermore, this result confirms previous reports stating that the Radviz technique is useful for highly dimensional datasets [12]. Figure 2(c) shows the visual quality R of the Radviz projections of datasets containing from 5 to 100 different classes. Similarly to the number of instances, we observe that as the number of classes increases, the quality of the projected data decreases logarithmically. Figure 2(d) that the maximum value of R is linearly reduced as the percentage of overlap increases (a bad Radviz projection may be bad because of the technique itself or may be bad due to the fact that the initial dataset is hardly separable). Thus, the computation of the DB index for the initial dataset can give us an insight on how well the Radviz visualization can do. Moreover, we want to understand whether the formalization of the optimization function that explores the DA has an impact in the way the optimal solution is obtained. The optimization function associates a numerical value to each of the DAs. Our Independent function (*indep*) looks for the highest value (maximization problem), and our dependent function (*dep*) looks for the smallest value (minimization problem). In order to understand the *quality* of the search space, we evaluate its non-convexity. The non-convexity of the search space gives a measure of the probability that the optimization function will fall into a local minima. The smaller the non-convexity of the search space, the higher the probability of a local minima (or maxima) being a global minima (or maxima). We calculate the non-convexity of the search spaces using the Haussdorf distance as $\lambda(A) = sup_{x \in co\ A}\ inf_{y \in A} \parallel x - y \parallel$ where A is the set of points in the space search and $co\ A$ represents its convex hull. We compute the non-convexity for the DA formalizations presented in Section 4: Independent function *indep*, Radviz-dependent function *dep* and the binary neighborhood matrix initially described by Ankerst *et al.* [1] (referred from now on as *Original*). As we can

observe in Figure 2(e), the *Original* function presented in [1] has a search area that is much less convex than the other two optimization functions for all possible combination of parameters and datasets. Still, such gap grows as the number of dimensions increases. These results indicate a higher probability of finding a global minimum (or maximum) when using the formalizations proposed in this paper.From previous analysis, the metric does not seem to impact the visual quality of the projections in terms of number of instances, classes, dimensions or percentage of overlap. In fact, we observe similar R values for both KL and COS across all the analysis. However, we want to understand whether the metric has an impact in the way the search space is explored *i.e.,* whether the selection of a metric can help decrease the chances of the optimization algorithm falling into a local minima (or maxima). For that purpose, we compute the non-convexity of the search spaces explored when using any combination of parameters.Figure 2(f) shows the trend between the non-convexity values of all combinations of parameters for each KL and COS metric. We can observe that the COS metric has smaller non-convexity values than KL. Hence, although in principle both metrics can potentially find solutions with similar visual quality R, the COS metric decreases the chances of the optimization function falling into local minima (or maxima), thus increasing the probability of finding a better DA.



(a) Visual quality: instances (b) Visual quality: dimensions (c) Visual quality: classes

(d) Visual quality: overlap (e) Non-convexity (opt.functions) (f) Non-convexity (metrics)

Fig. 2. Impact of the parameters in the visual quality of the projections: (a) instances, (b) dimensions, (c) classes, and (d) overlap; analysis of non-convexity according to (e) optimization functions and (f) metrics with both real and synthetic data.

6 Conclusions

Radviz (and radial visualizations) is one of the most common techniques to help in the process of detecting patterns when visualizing high dimensional data. One of the main problems of these techniques is that the usefulness of the projections highly depends on the dimension arrangement (DA), which is a NP-complete problem. In this paper, we have presented two novel variants for the formalization of the DA problem showing that they allow to explore a search space whose non-convexity makes it more probable to find the desired global maxima (minima). Then, we have presented a technique to automatically evaluate the visual usefulness of a projection by means of the Davies-Bouldin index, studying the relationships and the impact of various metrics and parameters in the quality of the visualization.

References

1. M. Ankerst, S. Berchtold, and D. A. Keim. Similarity clustering of dimensions for an enhanced visualization of multidimensional data. In *INFOVIS*, 1998.
2. M. Ankerst, D. A. Keim, and H.-P. Kriegel. Circle segments: A technique for visually exploring large multidimensional data sets. In *Visualization*, 1996.
3. D. L. Davies and D. W. Bouldin. A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-1(2):224–227, 1979.
4. P. Hoffman, G. Grinstein, K. Marx, I. Grosse, and E. Stanley. Dna visual and analytic data mining. In *VIS*, 1997.
5. P. Hoffman, G. Grinstein, and D. Pinkney. Dimensional anchors: a graphic primitive for multidimensional multivariate information visualizations. In *NPIVM*, 1999.
6. A. Inselberg and B. Dimsdale. Parallel coordinates: a tool for visualizing multidimensional geometry. In *VIS*, pages 361–378, Los Alamitos, CA, USA, 1990.
7. E. Kandogan. Star coordinates: A multi-dimensional visualization technique with uniform treatment of dimensions. In *IEEE Information Visualization Symp.*, 2000.
8. F. Kovács and R. Iváncsy. Cluster validity measurement for arbitrary shaped clusters. In *AIKED*, Wisconsin,USA, 2006.
9. G. Leban, B. Zupan, G. Vidmar, and I. Bratko. Vizrank: Data visualization guided by machine learning. *Data Min. Knowl. Discov.*, 13(2):119–136, 2006.
10. W. Peng, M. O. Ward, and E. A. Rundensteiner. Clutter reduction in multidimensional data visualization using dimension reordering. *InfoVis*, 0, 2004.
11. Y. A. R. F. Y. B.-Y. D. L. O. L. Y. Schler. Circle graphs: New visualization tools for text-mining. In *PKDD*, 1999.
12. J. Sharko, G. Grinstein, and K. A. Marx. Vectorized radviz and its application to multiple cluster datasets. *Visualization and Computer Graphics, IEEE*, 2008.
13. J. Yang, W. Peng, M. O. Ward, and E. A. Rundensteiner. Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In *Proc. IEEE Symposium on Information Visualization*, 2003.
14. J. Yang, M. O. Ward, and E. Rundensteiner. Visual hierarchical dimension reduction for exploration of high dimensional datasets, 2003.
15. H. Zhu. On information and sufficiency, 1997.