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 [11] 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

M.J. Zaki et al. (Eds.): PAKDD 2010, Part II, LNAI 6119, pp. 125–132, 2010.

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 [14]. 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.* [12] 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* [13] 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