Multidimensional clusters in RadViz

LENKA NOVÁKOVÁ, OLGA ŠTĚPÁNKOVÁ
Department of Cybernetics, Faculty of Electrical Engineering
Czech technical University in Prague
Technicka 2, Prague 6
CZECH REPUBLIC

Abstract: - The paper reviews those properties of RadViz visualization method [2] mapping data from n dimensional space into the plane which are important for identification of clusters in the multidimensional original data. It uses 2 characteristic examples of datasets which clearly point to a certain drawback of the original RadViz mapping. The identified problem can be resolved using 2 minor modifications of the RadViz algorithm which are suggested in the last section of this contribution.

Key-Words: - Clustering, Data Visualization, Data mining, RadViz

1 Introduction

Data-mining (DM) is a typical interdisciplinary activity, which requires close cooperation between two types of experts: those from the problem domain (e.g. medicine, banking or marketing) and those from the DM or machine learning community. Domain experts do not have to be present during all the data mining process, which can be rather time consuming. Often, there is no single way how to proceed in mining given dataset. On the contrary, there is a number of different approaches which can be applied to given data as well as number of questions which can be asked during the data mining process and the domain experts have crucial role in making the appropriate choice. They are supposed to help in focusing attention towards the most promising direction while taking into account the domain knowledge and the results, which have been obtained already. This type of involvement of the domain experts is most welcome because it ensures that the obtained results are really interesting and useful. On the other hand one has to take into account that the domain experts are most busy persons who never have enough of time. If too much time is required from the domain expert in a certain DM phase it can easily happen that this person will no more be willing to answer some urgent questions in further phases. That is why it is important to have quick means how to describe achieved results as well as to explain why we are asking some questions and what could be outcome of certain decisions. Data visualization can play a decisive role in this process because human mind excels in prompt interpretation of visual information. It is not surprising that Soukup and Davidson [4] claim that ”Visualization is a key in assisting business and data analysis to discover new patterns and trends in their business data sets.”

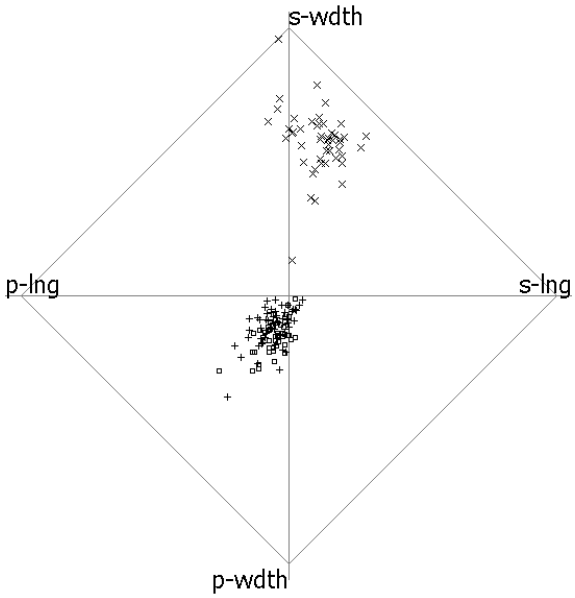

Fig.1 Visualization of the Iris dataset in RadViz [6]

One of the outcomes of the DM can be identification of characteristic clusters in which significant part of data appears or which exhibit some properties, which are different from those observed in the remaining part of the considered dataset. Such clusters can be easily identified if it is possible to see position of the considered data in the space. This is simple and natural if data is described using 2 or 3 properties (attributes): in such a case we can position each data point in 2D or 3D and look for places with higher density. RadViz [2] has been designed to map data described by 3 or more attributes (multidimensional data) to a planar picture. This contribution points to the fact that there are some significant clusters in mul-

tidimensional space, which cannot be identified in the corresponding RadViz picture. We suggest certain enhancement or modifications of RadViz which make up for this drawback.

Our paper is structured as follows: Section 2 reviews the definition of RadViz mapping and Section 3 provides a simple example of two 3D clusters which disappear in the corresponding RadViz picture. In Section 4, we derive conditions which ensure that the distance of 2 points in multidimensional space is closely related to the distance in RadViz picture. The Section 5 suggests minor modifications to RadViz which make the resulting method more useful for depicting multidimensional clusters - example from the Section 2 is used to illustrate this claim. The final Section 7 draws Conclusions from our observations.

2 RadViz

RadViz (Radial Coordinate visualization) is a visualization method, which uses the Hooke's law from physics for mapping a set of n-dimensional points into a plane. It offers a unique method which can help to identify relations among data. Its main advantage is that it needs no projections and provides a global view on the multidimensional data.

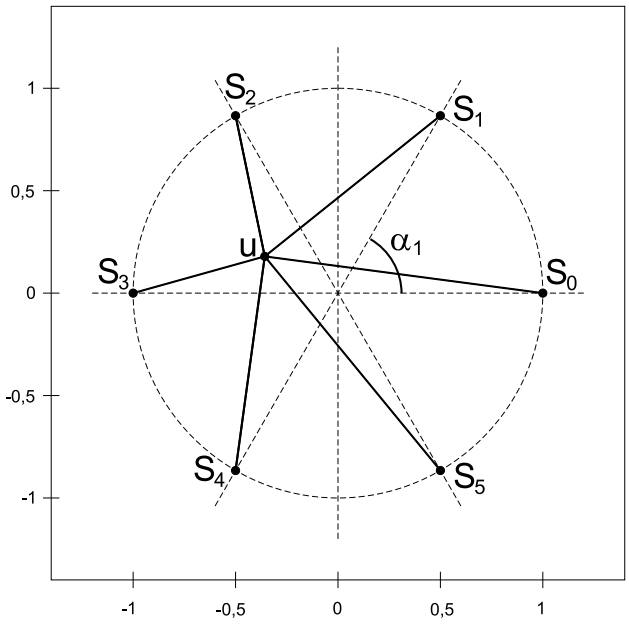

Fig.2 Definition of RadViz mapping

2.1 Radviz method theory

Each RadViz mapping of points from n dimensional space into a plane is uniquely defined by position of the corresponding n anchors (points S_j in the Fig. 2), which are placed in a single plane. The anchors are most often situated around a circle, but this is not necessary. Moreover, it is supposed that each anchor holds its own virtual spring of variable stiffness and all the loose ends of the springs are bound together. Let us consider a point $[y_1, \ldots, y_n]$ from n dimensional space. This point is mapped into a single point u in the plane of anchors as follows: for each anchor j the stiffness of its spring is set to y_j and the Hooke's law of mechanics is used to find the point u, where all the spring forces reach equilibrium (they sum to 0). The position of $u = [u_1, u_2]$ is given by the equations (1) or (2).

$$\begin{aligned} \sum_{j=1}^{n} (\overrightarrow{S}_j - \overrightarrow{u}) y_j &= 0 \\ \sum_{j=1}^{n} \overrightarrow{S}_j y_j &= \overrightarrow{u} \sum_{j=1}^{n} y_j \qquad (1) \\ \overrightarrow{u} &= \frac{\sum_{j=1}^{n} \overrightarrow{S}_j y_j}{\sum_{j=1}^{n} y_j} \qquad (2) \\ u_1 &= \frac{\sum_{j=1}^{n} y_j \cos(\alpha_j)}{\sum_{j=1}^{n} y_j} \\ u_2 &= \frac{\sum_{j=1}^{n} y_j \sin(\alpha_j)}{\sum_{j=1}^{n} y_j} \end{aligned}$$

2.2 Radviz algorithm

The visualisation algorithm using the RadViz method proceeds in the following steps:

1. Normalize the data to interval $\langle 0, 1 \rangle$,

$$\overline{x}_{ij} = \frac{x_{ij} - min_j}{max_j - min_j}$$

2. Place the dimensional anchors

3. Calculate the point where to place each record and draw it.

$$y_i = \sum_{j=1}^{n} \overline{x}_{ij}, \; \overrightarrow{u}_i = \frac{\sum_{j=1}^{n} \overrightarrow{S}_j \overline{x}_{ij}}{y_i}$$

3 RadViz and its basic properties

RadViz mapping proves useful to point to existence of natural clusters in some datasets, see for example RadViz image of Iris data - Fig. 1. But interpretation of the resulting RadViz image is not as innocent as it could look like. It is important to keep in mind that the used mapping is "many-to-one", what is precisely

expressed in the *Observation 1* (a simple consequence of (2)): All points from the n-dimensional space, which lay on a single line crosscutting the $[0, ..., 0]$, are mapped into a single point in the RadViz plane. Due to this property it can easily happen that some clusters, which are clearly separated in n-dimensional space, disappear in the RadViz image.

3.1 Some examples of superimposed clusters

For simplicity, let us illustrate the upper claim using data in 3D. Obviously, the idea of the presented examples can be easily reused in any higher dimension. Let us consider artificial datasets generated by a system PreDo [5], which has been designed at CTU to support experiments in machine learning domain.

Example 1. Let the dataset consist of two clusters of points depicted in the Fig. 3:

- all data in the first cluster are part of the intersection between the cube $\langle 0, 1\rangle^3$ and a sphere of radius 0.2 with center in $[0, 0, 0]$,
- data in the second cluster belong to the intersection of the same cube with a thick envelope of a sphere with radius between 0.8 and 0.9 which has the center in the center of coordinate system $[0, 0, 0]$ again.

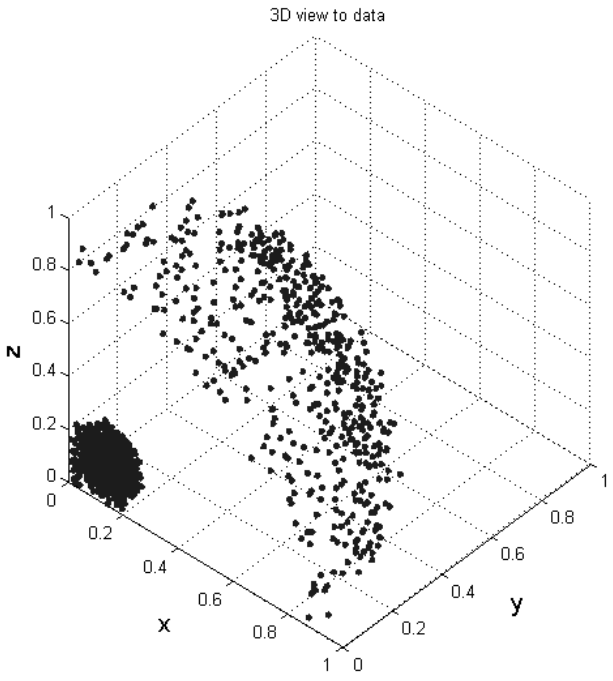

Fig.3 Example 1: Data in 3D

Obviously both clusters are clearly separated in the original dataset, but no hint of separation can be observed in the corresponding RadViz image - see Fig. 4. The datapoints from the small sphere appear all over the resulting planar picture.

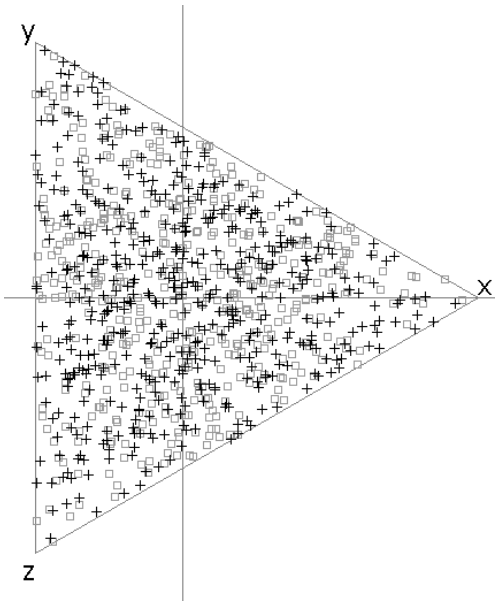

Fig.4 RadViz image of Example 1

Example 2. Let the dataset consist of two small spheres, which can be seen in the Fig. 5.

- all data in the light cluster belong to the sphere with the center [0.1,0.1,0.1] and radius 0.1,
- all data in the dart cluster belong to the sphere with the center [0.5,0.2,0.8] and radius 0.1.

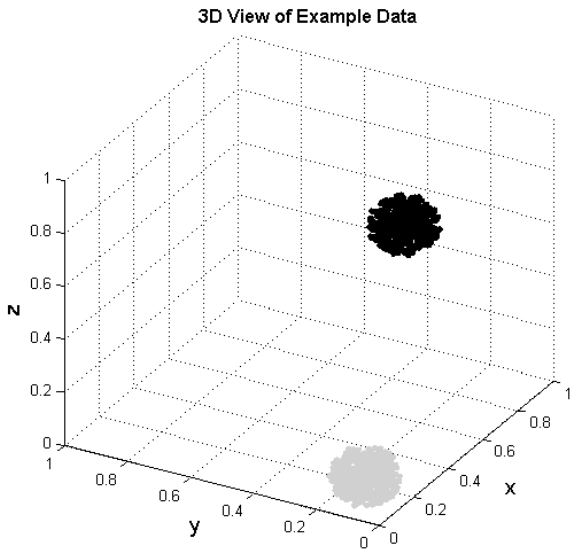

Fig.5 Example 2: Data in 3D

Here again one can see that the data belonging to the small dark sphere are totally lost among the images of points from the lighter sphere (which is close to the point $[0,0,0]$), see Fig. 6.

In the next section we try to specify conditions, which prevent this happening.

4 Estimate of distance in RadViz

We would like to find out under what conditions the distance between RadViz images of points, which are close to each other in n-dimensional space, remains small. In order to provide an estimate of distance between 2 points after RadViz mapping let us consider a point $a = [a_1, \ldots, a_n]$ inside the cube $\langle 0,1\rangle^n$ and a point $[a_1+\delta_1, \ldots, a_n+\delta_n]$ in its close neighborhood, ie. (it is supposed that for all dimensions j there holds $|\delta_j| < \delta \ll 1$ and $0 \leq a_n < 1$). Let us enumerate the distance of their images in both planar coordinates, namely $[d_1, d_2]$ see (3).

$$d_1 = \left| \frac{\sum\limits_j (a_j+\delta_j)\cos(\alpha_j)}{\sum\limits_j (a_j+\delta_j)} - \frac{\sum\limits_j (a_j)\cos(\alpha_j)}{\sum\limits_j (a_j)} \right| =$$

$$= \left| \frac{\sum\limits_j a_j \sum\limits_j \delta_j \cos\alpha_j - \sum\limits_j \delta_j \sum\limits_j a_j \cos\alpha_j}{\sum\limits_j (a_j+\delta_j) \sum\limits_j a_j} \right| \qquad (3)$$

Using the 2 straightforward facts :

- the upper estimate of the absolute value of a linear combination of elements is given by a sum of absolute values of these elements,
- $|cos(\alpha)|$ is ever smaller than 1,

we can reach the conclusion that (4) holds.

$$d_1 < \frac{2\sum\limits_j a_j \sum\limits_j \delta_j}{\sum\limits_j (a_j+\delta_j) \sum\limits_j a_j} = \frac{2\sum\limits_j \delta_j}{\sum\limits_j (a_j+\delta_j)} \qquad (4)$$

Now it is clear that the distance of the considered 2 points strongly depends on the position of a. If this point is very close to $[0, \ldots, 0]$, the images of a and $a+\delta$ can appear far apart. On the other hand, let us suppose that we can rely on a reasonable lower estimate of $|a_j+\delta_j|$ for all dimensions. If $a_j > 1/m$ where m is a fixed natural number bigger than 1, e.g. $m = 2$, we can be sure that for all j there holds $|a_j+\delta_j| > \frac{1}{3}$ and since $|\sum\limits_j (a_j+\delta_j)| > \frac{n}{3}$ we can consequently claim that (5) holds.

$$d_i < \frac{2\sum\limits_j |\delta_j|}{\sum\limits_j (a_j+\delta_j)} < \frac{2\cdot 3\delta}{n} = 6\delta \qquad (5)$$

Supported by the estimate (5) we can express the *Observation 2*: If the distance between 2 points, which both appear outside of the cube $\langle 0, 0.5\rangle^n$, is small then their RadViz images cannot be far away from each other. This means that small clusters in n-dimensional space remain small even in the RadViz image.

5 Conclusions

When using the RadViz method for data visualization one should not forget about both *Observations 1* and *2* mentioned in the Sections 3 and 4.

The problem depicted by the Example 2 from the Section 2 could be easily resolved if the normalization scale applied to the original n-dimensional data in the step 1 of 2.2 RadViz algorithm is changed - see Fig. 6, 7 and 8 .

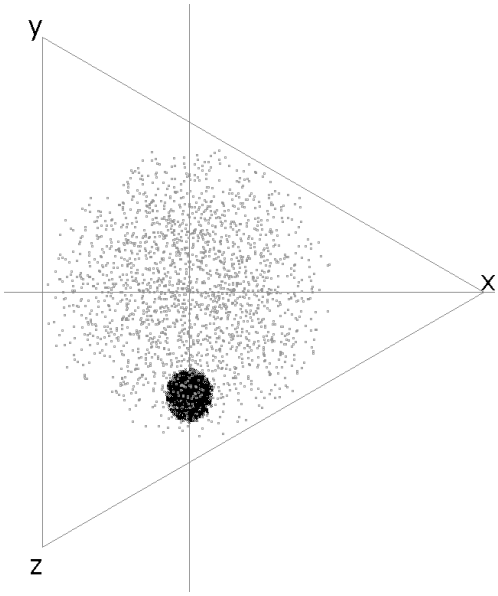

Fig.6 RadViz image for Example 2, normalization to the interval $\langle 0, 1\rangle$

The situation is not as simple for the data from the Example 1, where the size of one of the clusters is far from small. Even here the change of normalization scale helps - see Fig. 9. All over it the only way how both clusters of Example 1 can be clearly separated

(see Fig. 10) is by making use of a modification **RadVizS** [3] mapping the data points into 3D instead of to 2D. The mapping is defined as follows:

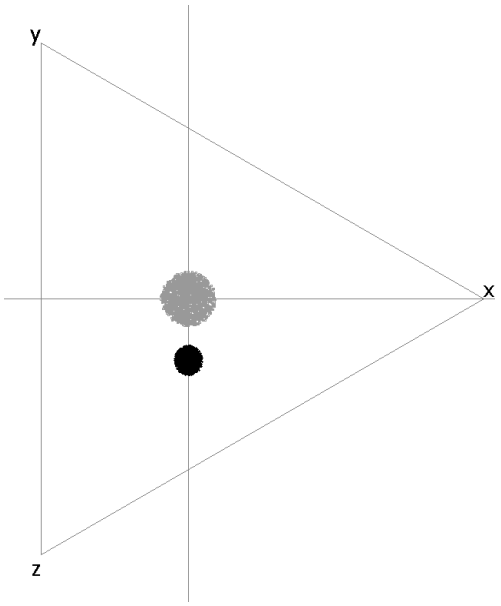

Fig.7 RadViz image for Example 2, normalization to the interval $\langle 0.25, 1\rangle$

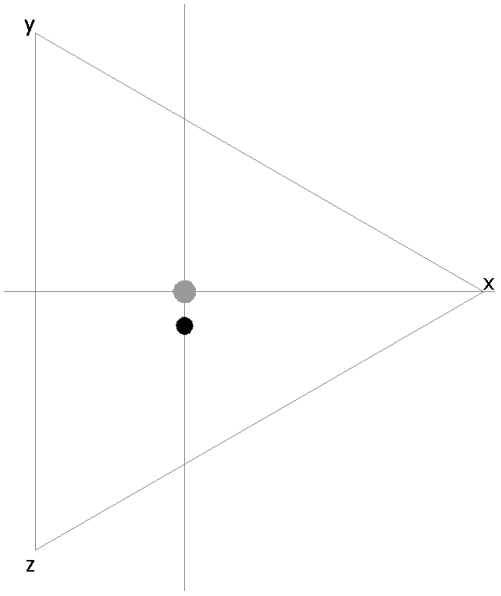

Fig.8 RadViz image for Example 2, normalization to the interval $\langle 0.5, 1\rangle$

Let the RadViz image of the point a from n dimensional space be $u = [u_1, u_2]$. RadVizS maps the point a to $u = [u_1, u_2, u_3]$, where u_3 is the Eucleidean distance of the point a from the coordinate origin $[0, \ldots, 0]$.

When using RadViz for data visualization, we recommend to change the normalization scale suggested above, i.e. to normalize the data into the interval $\langle 0.5, 1\rangle$ instead to the originally suggested transformation into $\langle 0, 1\rangle$. To get even more precise picture of the considered data it is useful to depict them using RadVizS.

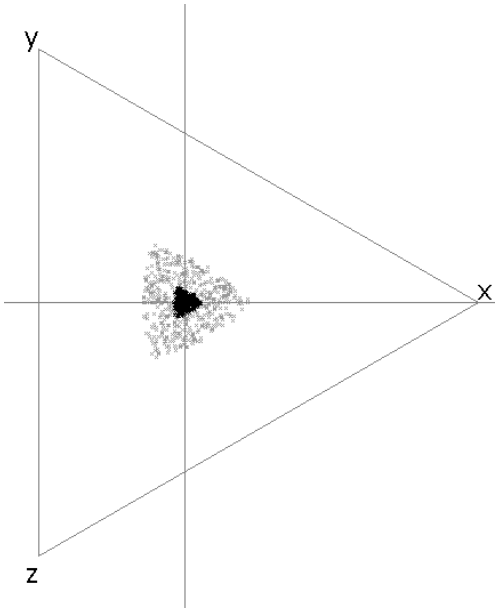

Fig.9 RadViz image for Example 1, normalization to the interval $\langle 0.5, 1\rangle$

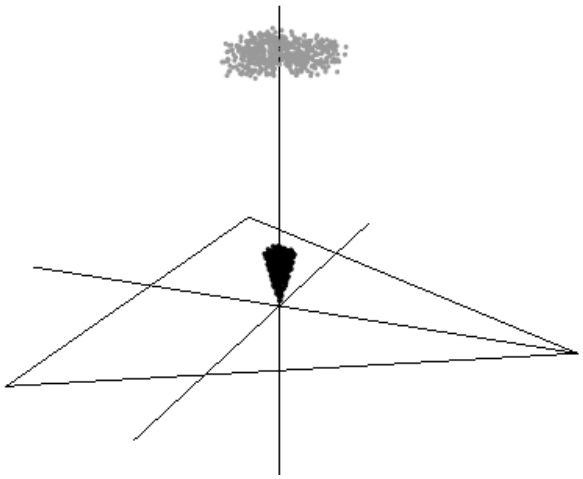

Fig.10 RadvizS for Example 1 with normalization to the interval $\langle 0.5, 1\rangle$

We extensively utilize all upper mentioned modifications of RadViz in analyzing real life data. This process is supported by the original system for data preprocessing SumatraTT [1], [7], which includes also our implementation of the RadViz method.

6 Acknowledgement

The presented research and development has been partially supported by the grant of the FRVS agency of the Czech Ministry of Education 832/2006 ”Innovation of the lectures in Artificial Intelligence”.

References:

[1] Aubrecht P., Kouba Z.:A Universal Data Preprocessing System. *Datakon 2003*, Brno, 2003

[2] Fayyad U., Grinstein G. G., Wierse A.: *Information Visualization in Data Mining and Knowledge Discovery.* Morgan Kaufmann Publishers, 2002.

[3] Novakova L., Stepankova O.: Visualization of Some Relational Patterns for DM, *Cybernetics and Systems 2006*, Vol.2, Vienna, 2006, pp.785-790

[4] Soukup T. and Davidson I.: *Visual Data Mining. Techniques and Tools for Data Vizualization and Mining.* John Wiley & Sons, Inc., 2002.

[5] Vejmelka M.: PreDO - Precisely Defined Objects, Research Report GL, Prague, 2002.

[6] Iris Dataset, UCI Machine Learning Repository, www.ics.uci.edu/ mlearn/MLRepository.html

[7] SumatraTT, WWW homepage, http://krizik.felk.cvut.cz/Sumatra