Density-Based Multidimensional Scaling

F. Rehm, F. Klawonn, and R. Kruse

Abstract Multidimensional scaling provides dimensionality reduction for high-dimensional data. Most of the available techniques try to preserve similarity in terms of distances between data objects. In this paper a new approach is proposed that extends the distance preserving aspect by means of density preservation. Combining both, the distance aspect and the density aspect, permits efficient multidimensional scaling solutions.

1 Introduction

Most branches of commerce, industry and research put great efforts in collecting data with the objective to describe and predict customer behaviour or both technical and natural phenomena. Besides the size of such data sets, data analysis becomes challenging due to a large number of attributes describing a data object. Visualization can facilitate the discovery of structures, patterns and relationships in data and exploratory visualization is an important component in hypothesis generation.

Multidimensional scaling (MDS) is a family of dimensionality reduction techniques that use optimization to preserve distance relationships between points in the multidimensional space in the two- or three-dimensional mapping required for effective visualization (Kruskal and Wish 1978). In the recent years much effort has been done to improve MDS regarding its computational complexity (Borg and Groenen 2005; Chalmers 1996; Morrison et al. 2003; Williams and Munzner 2004). Besides distance-based approaches also some techniques preserving angles between data objects have been applied successfully (Lesot et al. 2006; Rehm et al. 2006).

In this paper we present a new approach that extends conventional distance-based multidimensional scaling by a density preserving aspect. This permits to improve

F. Rehm(✉)
Institute of Flight Guidance, German Aerospace Center, Braunschweig, Germany,
E-mail: frank.rehm@dlr.de

A. Okada et al. (eds.), *Cooperation in Classification and Data Analysis*,
Studies in Classification, Data Analysis, and Knowledge Organizations.
DOI: 10.1007/978-3-642-00668-5_5,

the mapping of high-dimensional data for visualization purposes. The rest of the paper is organized as follows. In Sect. 2 we briefly review Sammon's mapping as a common representative of distance-based MDS. Section 3 describes the proposed method. Section 4 discusses results on benchmark examples. Finally we conclude with Sect. 5.

2 Sammon's Mapping

Sammon's mapping is a multidimensional scaling technique that estimates the coordinates of a set of objects $Y = \{y_1, \ldots, y_n\}$ in a feature space of specified (low) dimensionality that come from data $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^p$ trying to preserve the distances between pairs of objects. These distances are usually stored in a distance matrix

$$D^x = \left(d_{ij}^x\right), \quad d_{ij}^x = \left\|x_i - x_j\right\|, \quad i, j = 1, \ldots, n.$$

The estimation of the coordinates will be carried out under the constraint that the error between the distance matrix D^x of the data set and the distance matrix $D^y = (d_{ij}^y)$, $d_{ij}^y = \|y_i - y_j\|$, $i, j = 1, \ldots, n$ of the corresponding transformed data set will be minimized.

Different error measures to be minimized were proposed, e.g., the absolute error that considers non-weighted differences between original distances and distances in the target space, the relative error that takes relative distances into account or a combination of both. The Sammon's mapping error measure

$$E_{\text{sammon}} = \frac{1}{\sum_{i=1}^{n} \sum_{j=i+1}^{n} d_{ij}^x} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{\left(d_{ij}^y - d_{ij}^x\right)^2}{d_{ij}^x} \tag{1}$$

describes the absolute and the relative quadratic error. To determine the transformed data set Y by means of minimizing error E_{sammon} a gradient descent method can be used. By means of this iterative method, the parameters y_l to be optimized, will be updated during each step proportional to the gradient of the error function E. Calculating the gradient of the error function leads to

$$\frac{\partial E_{\text{sammon}}}{\partial y_l} = \frac{2}{\sum_{i=1}^{n} \sum_{j=i+1}^{n} d_{ij}^x} \sum_{j \neq l} \frac{d_{lj}^y - d_{lj}^x}{d_{lj}^x} \frac{y_l - y_j}{d_{lj}^y}. \tag{2}$$

After random initialization for each projected feature vector y_l a gradient descent is carried out and the distances d_{ij}^y as well as the gradients $(\partial d_{ij}^y / \partial y_l)$ will be recalculated again. The algorithm terminates when E_{sammon} becomes smaller than a certain threshold.

3 Density-Based Mappings

The concept of density-based visualization is to map density distributions of high-dimensional data into low-dimensional feature spaces. Density variations often indicate the existence of clusters which, commonly, are of concern in the field of data mining. Thus, projecting these density distributions to a visually interpretable display may help to identify interesting patterns in the data.

In the following we formalize the problem of density preservation by means of an objective function that can be minimized through a gradient descent technique. For each data object in the original data space a multivariate Gaussian distribution is defined that represents a data point's potential energy. When adding those single potentials we get a sort of multidimensional potential mountains. Summits of the mountains can be found where many data objects are located. Accordingly, valleys can be found in areas of low data density.

Similarly, one can reproduce the mountains in the low-dimensional feature space (usually two or three dimensions). For this purpose each data object of the original space will be placed in the projection space. Over every single data point a potential (in form of a two- or three-dimensional Gaussian distribution) will be applied. The criterion for the mapping is that the potentials in the original space coincide as good as possible with the potentials at the corresponding points in the target space.

Given the data set $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^p$ we seek for the mapped data set $Y = \{y_1, \ldots, y_n\} \subset \mathbb{R}^k$ with $k = 2$ or $k = 3$ with the following potential for x_i:

$$f_i(x) = \frac{1}{c} \exp\left[-\frac{1}{2} \sum_{t=1}^{p} \left(\frac{x^{(t)} - x_i^{(t)}}{\sigma} \right)^2 \right] \tag{3}$$

with

$$c = \frac{1}{\sigma^p \sqrt{(2\pi)^p}}.$$

By $x^{(t)}$ and $x_i^{(t)}$ we denote the t-th attribute of data object x and x_i, respectively. Function f_i simply describes the density of a p-dimensional Gaussian distribution with mean value x_i and variance σ^2 in each dimension. The parameter σ must be fixed for the entire procedure. If σ is rather small, then the potentials do rarely overlap. For very large σ the potential landscape will be blurred completely with little variance in height.

Therefore, it is useful to define σ according to the diameter d of the data space, the average distance between data points, the number n of data objects and the dimensionality p. A straight forward approach would be to assume that the data is uniformly distributed in a hyper-cube or hyper-sphere. In this case the potentials would have approximately the same height. Of course, this assumption is fairly theoretical. In practice mountains will be formed due to the heterogeneous structure of the data. However, under this assumption the average density can be computed and the potentials on and between data points can be determined. The larger the

variance σ^2, the smaller the difference in the potentials. For small data sets the density is low and therefore a larger σ should be chosen.

Similar to Sammon's mapping we seek the projected data points $Y = \{y_1, \ldots, y_n\} \subset \mathbb{R}^k$. Over each data point we apply a potential (in this case a k-dimensional Gaussian distribution) as for the original space:

$$g_i(y) = \frac{1}{\tilde{c}} \exp\left[-\frac{1}{2}\sum_{t=1}^{k}\left(\frac{y^{(t)} - y_i^{(t)}}{\tilde{\sigma}}\right)^2\right] \tag{4}$$

with

$$\tilde{c} = \frac{1}{\tilde{\sigma}^k\sqrt{(2\pi)^k}}.$$

Then the objective is to place the feature vectors such that the potentials coincide at least in these points with those in the original space. Note, $\tilde{\sigma}$ should be chosen similarly to σ. In the ideal case we have approximately the same diameter d in the target space, too. However, the area (or the volume) of the target space will be much smaller compared to the hyper volume of the original space ($k \ll p$). This means that the density in the target space is also higher for the same size of the data set. Thus, $\tilde{\sigma}$ should be chosen smaller than σ. Still the potentials in the target space might not match the potentials in the original space yet. It should be assured that the maximum height of the single potentials in the original space and in the target space match, i.e., the respective maxima of the Gaussian distributions should be:

$$f_i(x_i) \approx g_i(y_i).$$

Since normally this will not be the case we introduce a constant a:

$$af_i(x_i) = g_i(y_i)$$

which can be derived from (3) and (4):

$$a = \frac{\sigma^p}{\tilde{\sigma}^k}\sqrt{(2\pi)^{p-k}}.$$

Now we can formulate our objective function. The summarized modified potential in the original space at x_i is

$$\sum_{j=1}^{n} af_j(x_i)$$

and in the target space at y_i

$$\sum_{j=1}^{n} g_j(y_i).$$

In the ideal case, both potentials should be equal. Hence, we define the objective function as follows:

$$E_{\text{density}} = \sum_{i=1}^{n} \left[\sum_{j=1}^{n} g_j(y_i) - \sum_{j=1}^{n} a f_j(x_i) \right]^2 = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{n} \left[g_j(y_i) - a f_j(x_i) \right] \right\}^2. \tag{5}$$

Now, we only have to determine the gradient for each component s:

$$\frac{\partial E_{\text{density}}}{\partial y_{ls}} = 2 \sum_{i=1}^{n} \sum_{j=1}^{n} \left[g_j(y_i) - a f_j(x_i) \right] \frac{\partial}{\partial y_{ls}} g_j(y_i). \tag{6}$$

$\frac{\partial}{\partial y_{ls}} g_j(y_i)$ is only zero when we have $l = i$ or $l = j$. For both cases we derive from (6):

$$\frac{\partial}{\partial y_{ls}} g_l(y_i) = \frac{1}{\tilde{c}} \exp\left[-\frac{1}{2} \sum_{t=1}^{k} \left(\frac{y_i^{(t)} - y_l^{(t)}}{\tilde{\sigma}} \right)^2 \right] \frac{y_i^{(s)} - y_l^{(s)}}{\tilde{\sigma}}, \tag{7}$$

$$\frac{\partial}{\partial y_{ls}} g_j(y_l) = -\frac{1}{\tilde{c}} \exp\left[-\frac{1}{2} \sum_{t=1}^{k} \left(\frac{y_l^{(t)} - y_j^{(t)}}{\tilde{\sigma}} \right)^2 \right] \frac{y_l^{(s)} - y_j^{(s)}}{\tilde{\sigma}}. \tag{8}$$

It can be easily seen that for $i = j = l$ we have $\frac{\partial}{\partial y_{ls}} g_j(y_i) = 0$. Finally we obtain for the gradient:

$$\begin{aligned} \frac{\partial E_{\text{density}}}{\partial y_{ls}} = \frac{2}{\tilde{c}} \sum_{i=1}^{n} \Bigg\{ & \left[g_l(y_i) - a f_l(x_i) \right] \\ & \times \exp\left[-\frac{1}{2} \sum_{t=1}^{k} \left(\frac{y_i^{(t)} - y_l^{(t)}}{\tilde{\sigma}} \right)^2 \right] \frac{y_i^{(s)} - y_l^{(s)}}{\tilde{\sigma}} \\ & - \left[g_i(y_l) - a f_i(x_l) \right] \\ & \times \exp\left[-\frac{1}{2} \sum_{t=1}^{k} \left(\frac{y_l^{(t)} - y_i^{(t)}}{\tilde{\sigma}} \right)^2 \right] \frac{y_l^{(s)} - y_i^{(s)}}{\tilde{\sigma}} \Bigg\}. \end{aligned} \tag{9}$$

Combining the Sammon gradient E_{sammon} and the density gradient E_{density} through linear combination we finally obtain

$$E = \alpha \frac{\partial E_{\text{sammon}}}{\partial y_l} + \beta \frac{\partial E_{\text{density}}}{\partial y_l}. \tag{10}$$

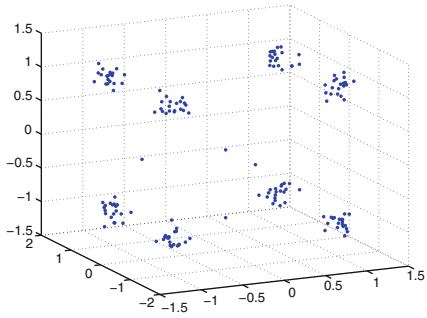

Fig. 1 Cube data set

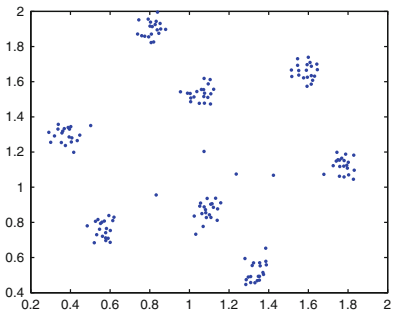

Fig. 2 Sammon's mapping of the Cube data set

The parameters $\alpha \geq 0$ and $\beta \geq 0$ can be considered as learning rates or weights to control the impact of the respective mapping strategy. Thus, higher weights α for the Sammon gradient favour distance-based mappings and larger values β for the density gradient favour the density approach.

4 Results

In this section we will discuss some results of the proposed technique on some benchmark examples. The first data set, the Cube data set (see Fig. 1), is about a synthetic data set, where data points scatter around the corners of an imaginary three-dimensional cube. Thus, the Cube data set contains eight well separated clusters. The second data set, the Wine data set (Forina et al. 1988), results from a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

Figure 2 shows a Sammon's mapping of the Cube data set. The eight data clusters are well reflected in the mapping. The transformation with the density-based approach, setting $\alpha = 0$ and thusly optimizing the density aspect exclusively, leads to the mapping visualized in Fig. 3. It is surprising that already the density aspect in the optimization is sufficient in this example to reflect the structure of the data set. Applying a linear combination of both, the Sammon gradient and the density

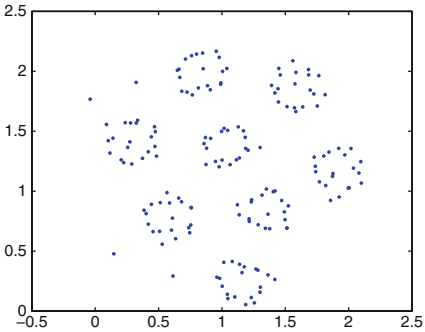

Fig. 3 Density-based mapping of the Cube data set

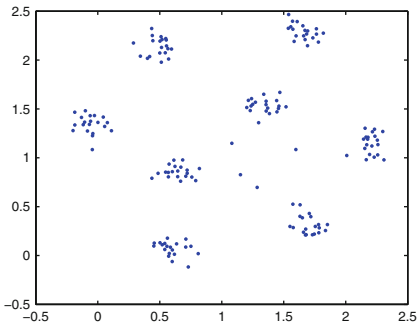

Fig. 4 Mapping of the Cube data set (distance-based and density-based)

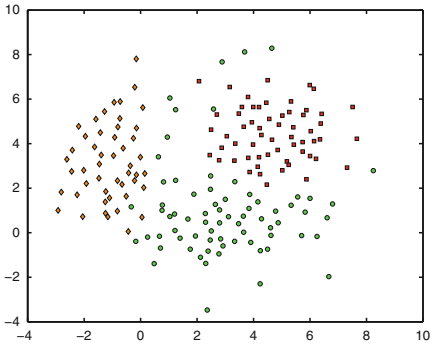

Fig. 5 Sammon's mapping of the Wine data set

gradient, we obtain the mapping depicted in Fig. 4. Whereas the distance-based approach seams to favour the preservation of the inter-cluster structure, the linear combination of distance and density aspects gives a better overall impression of the data set.

Figures 5 and 6 show transformations of the Wine data set with Sammon's mapping and with the density-based approach, respectively. Both transformations show similar characteristics.

Based on the empirical tests we cannot constitute that the density-based approach is superior to the distance-based approach. Indeed, the computational complexity per iteration of the density-based approach is rather higher since the density gradient

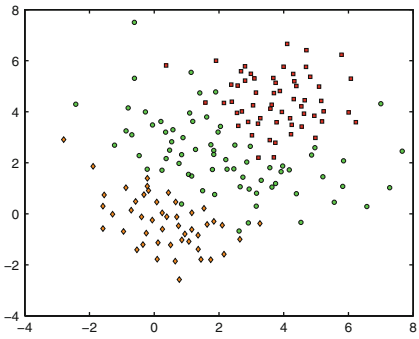

Fig. 6 Density-based mapping of the Wine data set

has to be computed additionally. Our tests have shown that the number of iterations can be reduced with density preservation.

5 Conclusions

In this paper we have presented a new approach to visualize high-dimensional data. Density-based multidimensional scaling considers not only the distance aspect as it is usual but also density aspects of a data set. We could show that our approach is promising and leads to comparable results as conventional MDS and can lead to better results in combination with MDS. Future work should focus on further tests on complex data sets to prove stability and convergence.

References

BORG, I. and GROENEN, P. (2005): *Modern Multidimensional Scaling: Theory and Applications*. Springer, Berlin.

CHALMERS, M. (1996): A Linear Iteration Time Layout Algorithm for Visualising High-Dimensional Data. In *Proceedings of IEEE Visualization 1996*, San Francisco, CA, 127–132.

FORINA, M., LEARDI, R., ARMANINO, C. and LANTERI, S. (1988): *PARVUS: An Extendable Package of Programs for Data Exploration, Classification and Correlation.* Elsevier, Amsterdam.

KRUSKAL, J.B. and WISH, M. (1978): *Multidimensional Scaling.* Sage, Beverly Hills.

LESOT, M.J., REHM, F., KLAWONN, F. and KRUSE, R. (2006): Prediction of Aircraft Flight Duration. In *Proceedings of the 11th IFAC Symposium on Control in Transportation Systems,* Delft, 107–112.

MORRISON, A., ROSS, G. and CHALMERS, M. (2003): Fast Multidimensional Scaling through Sampling, Springs and Interpolation. *Information Visualization*, 2, 68–77.

REHM, F., KLAWONN, F. and KRUSE, R. (2006): POLARMAP – Efficient Visualisation of High Dimensional Data. In *IEEE Proceedings of the 10th International Conference on Information Visualisation*, London, 731–740.

WILLIAMS, M. and MUNZNER, T. (2004): Steerable, Progressive Multidimensional Scaling. In *Proceedings of the 10th IEEE Symposium on Information Visualization,* Austin, TX, 57–64.