INTERACTIVE DATA EXPLORATION USING MDS MAPPING

Antoine Naud and Włodzisław Duch[1]
Department of Computer Methods
Nicolaus Copernicus University
ul. Grudziądzka 5, 87-100 Toruń, Poland

Abstract:

Interactive exploratory data analysis can be realised by using dimensionality reduction techniques integrated in data visualization software. This work presents an adaptation of one multidimensional scaling algorithm to provide it with generalization capability, allowing the display of new data on an existing mapping. The ensuing *relative mapping* is used to help the understanding of classification results.

Keywords: Data visualization, Exploratory Data Analysis, multidimensional scaling.

1 INTRODUCTION

In many applications understanding of the data is of primary importance. Thus usual approach to data analysis in machine learning community is based on logical rules and the lack of comprehensibility is perceived as the biggest drawback of neural network and pattern recognition methods [4]. Interpretation of logical rules is not so straightforward as it may seem since rules may not be stable and logical true-false answers near decision borders of the classifier are misleading. If the decision borders are complex, a large number of crisp rules are needed for proper approximation. Fuzzy rules may be more appropriate in such case, although they also do not provide optimal decision borders.

An alternative way to understand the structure of the data is based on visualization. New case, given for classification, should be viewed in relation to known cases stored in the training database and in relation to the decision borders provided by the classifier used. Visualization may always be used to understand decision borders of any classifier, and in some cases may provide more information than logical rules. Unfortunately our spatial imagination is restricted to 3 dimensions and thus we cannot directly view relations in highly dimensional spaces. Self-organizing topographic feature maps (SOM) [8], known also as the Kohonen networks, were design as a visualization and classification tool. They have found many applications as systems capable of unsupervised learning and visualization of data. Most classification problems belong to the supervised learning class. SOM networks appear to be among the worst classifiers in such tasks [12]. They also do not seem to provide good maps that do preserve correct topographical relations among multidimensional objects [6, 7]. Quantitative measures for the distortions of original data topography by SOM mapping have been introduced in [3]. The same idea has already been published in psychometric journal by Kruskal

[1] E-mail address: {naud, duch}@phys.uni.torun.pl

[10] and is known as “the multidimensional scaling" (MDS), and in engineering journal [14], where it is known as “the Sammon mapping".

A quantitative measure of topographical distortion based on differences between distances of objects in the original space and distances between projections of these objects in the low-dimensional space is introduced and minimized in MDS methods. In this paper we will show how MDS maps displaying classification borders may be formed. In the next section some information about the software allowing for the interactive data exploration is given, in the third section a method to place a single new object on existing map is described, and in the fourth section sensitivity of such maps to initial conditions and computational experiments are described. A short discussion concludes this paper.

2 INTERACTIVE DATA EXPLORATION USING MDS MAPPING

The TCM (Topographically Correct Mapping) interactive software developed by us performs a mapping of multivariate data from a high-dimensional space (D-dimensional space, $D \gg 3$) to data points in a lower d-dimensional space ($d \ll D$). Usually $d = 2$ in order to allow visualization by scatterplots. This dimensionality reduction, obtained by a multidimensional scaling (MDS) procedure, is such that similar (or close, in the sense of the Euclidean distance in the D-dimensional space) multivariate objects or cases are mapped on representative points close to each other in the d-dimensional representation space. The mapping should preserve topography of data vectors. Linear projections are not able to preserve topography; such mappings have to be non-linear. Topography preservation may be unreachable in the whole feature space if the analyzed multivariate data is intrinsically high dimensional and cannot be imbedded in a lower dimensional space without distortion. Therefore MDS mappings may be confusing and may give incorrect ideas about data structure. When zooming on a small neighborhood of a particular vector, distortions of such mapping should become smaller. Therefore the software should be interactive, allowing for such zooming.

The MDS algorithm used here (Kruskal’s nonmetric scaling [10]) requires only a dissimilarity or distance matrix of the multivariate objects. Preservation of data topography allows the display of the data structure, the clusters, outliers or class overlapping. The program should be treated as an exploratory data analysis tool and should have a Graphical User Interface to allow interactive exploration. Perhaps the best known software for interactive data exploration is the XGobi [15], allowing to see 3-dimensional projections from different perspective, but without preservation of multidimensional topographical relations. The minimization involved in TCM makes it difficult to provide real-time tour in the multidimensional space, as it is done with the Xgobi projections. In TCM the initial MDS algorithm has been adapted to perform MDS mapping of new points on a set of points previously mapped, i.e. to perform a *relative mapping*. The user may select a subset of a mapped dataset using the mouse and to create a new map, zooming on this subset only. The value of topographical distortion measure should decrease during zooming and allows an evaluation of the confidence one may have in correct representation of the original data topography.

3 RELATIVE MDS MAPPING

In classification tasks, it can be interesting to see where a new data points "falls" among known cases, and discover the class topology of its neighboring known cases (classified and labeled), to get an insight on how a classifier would classify this new data. The realization of this purpose gives rise to the need for a method that allows the mapping of one new point on a set of data points previously mapped, using a topology-preserving mapping. MDS is a topology preserving mapping, but it does not offer the possibility to project new points on an existing mapped set of points. To get a mapping presenting the previously mapped known points together with the new ones requires a complete re-run of the MDS algorithm on the new and the old data points. Let us denote by N_f the number of known data points, N_m the number of new data points, $N_t = N_m + N_f$ the total number of points considered during the mapping, $F = \{P_i, i = 1, N_f\}$ the set of known data points and by $M = \{P_i, i = 1, N_m\}$ the set of new data points. The computational cost involved can be reduced using the following scheme:

1. Map set F using normal MDS mapping,
2. Map set M relatively to the mapped set F using the “’relative” MDS mapping technique.

Relative MDS mapping differs from normal MDS by the fact that during the minimization of the topography distortion measure (which will be called “Stress" here) only the points from M are allowed to move while the points from F are kept fixed. This is achieved by modifying the Stress function so that it sums only over the distances that change during iterations, i.e. the distances between the fixed and the moving points, and the moving points interpoint distances. The original Stress function:

$$S(\mathbf{x}) = \sum_{ij}^{N_t} w_{ij} \cdot \left(\hat{d}_{ij} - d_{ij}\right)^2 \tag{1}$$

is redefined as:

$$S_r(\mathbf{x}) = \sum_{ij}^{N_m} w_{ij} \cdot \left(\hat{d}_{ij} - d_{ij}\right)^2 + \sum_{i=1}^{N_m} \sum_{j=N_m+1}^{N_t} w_{ij} \cdot \left(\hat{d}_{ij} - d_{ij}\right)^2 \tag{2}$$

Relative mapping can also be used to visualize large datasets, in which the high number of points makes MDS techniques unusable due to the exponentially increasing computation time. Various approaches have been proposed to tackle this problem, especially in the case of Sammon mapping. One category of such approaches, called in [2] the *frame method*, is to first select a subset of N_b points and to map it, and then to add sequentially the remaining points to the mapping. Relative mapping can be used to perform this second step, taking the points already mapped as fixed points, and the first point to be mapped as the moving point. In this way relative mapping offers an alternative to methods designed for the purpose of giving generalization capability to Sammon mapping such as the ANN Sammon mapping [9], Neuroscale [16], distance mapping [13] or incremental scaling [1].

4 SENSITIVITY OF MDS ALGORITHM TO INITIAL CONDITIONS AND COMPUTATIONAL EXPERIMENTS

MDS algorithm may be sensitive to initial conditions, making the interpretation of maps difficult. The configuration of representative points that is chosen to start the algorithm may have a strong influence on the final configuration. This is due to the fact that MDS relies on the minimization of a multivariate function by local minimization methods that get easily stucked in local minimums.

We have applied two strategies for algorithm initialization. The first one was to set the initial configuration randomly and map it. This is repeated a number of times (20 times is often enough) and only the resulting configuration with the lowest Stress measure is kept. The second strategy is to first perform a linear mapping using Principal Components Analysis (by a Singular Value Decomposition of the data matrix) and to use the two first principal components as the initial configuration.

Empirical experience on various datasets shows that the best Stress reached after random initialization is often slightly lower than the Stress resulting from the mapping initialized by PCA, but those differences are almost invisible when looking at the resulting configurations. In the case of a single point mapped relatively to the existing map, a better choice is to initialize the representative point by the triangulation method [11], leading in the D-dimensional space to the exact preservation of the distances to the d closest fixed points.

The dataset used in our visualization experiments contains results of 1611 psychometric evaluations (this is an extended version of the database used in [5]), with 13 attributes, belonging to one of 20 classes. Figure 1 shows zooming into the neighborhood of one particular point (marked as $\times$) from the database class 'organic_w'. The plots show the mappings of the 300, 200, 50, 40, 30 and 20 closest points to this chosen point $\times$ (the neighbors marked $\circ$ are from class 'norma_w', down triangles – class 'organic_w', $\#$ – class 'neurosis_w', up triangles – class 'psychopatia_w', $+$ – class 'schizofrenia_w', pentagons – class 'alcoholics' and rectangles – class 'narcotics_w').

5 DISCUSSION

Visualization may be sufficient to understand the data and may be the method of choice if the classification borders are complex. By separating classification from visualization, a better performance in both tasks may be achieved than that obtained with the SOM maps. Interpretation of these maps requires careful analysis of the amount of topographical distortions they introduce. Interactive zooming on interesting areas of the input space allows the exploration of the neighborhood of the case under inspection. Since there may be some uncertainty in the input one may generate a number (for example 100) of vectors drawn from a Gaussian distribution centered at this vector. The class of these additional points is determined using neural or other classification systems and corresponding points added (using relative mapping) to the map. Visual inspection of such map allows estimating the probability of misclassification, proportional to the number of points from alternative classes.

Visualization of the decision borders of different classifiers is another interesting possibility. In this case large number of additional points are produced, classified and mapped relatively to the true points from the database. All surface of the map becomes colored, showing smooth transitions between different classes. Areas that are near the classification borders will have mixed (dithered) colors, while areas containing well separated classes will have pure colors. In this way classification borders will be visible and the confidence in classification may be evaluated, depending on the distance of the given vector from the border. Topographically correct mapping may find many applications.

Acknowledgments: J. Gomuła and T. Kucharski are gratefully acknowledged for providing the psychometric data.

REFERENCES

[1] W. Basalaj. Incremental multidimensional scaling method for database visualization. In *Proceedings of Visual data Exploration and Analysis VI, SPIE*, volume 3643 (1999) 149-158.

[2] C. L. Chang and R. C. T. Lee. A heuristic relaxation method for nonlinear mapping in cluster analysis. *IEEE Transactions on Systems, Man, and Cybernetics* 3 (1973) 197–200

[3] W. Duch. Quantitative measures for the self-organizing topographic maps. *Open Systems & Information Dynamics* 2 (1995) 295-302

[4] W. Duch, N. Jankowski, K. Grąbczewski, and R. Adamczak. Optimization and interpretation of rule-based classifiers. In P. V. (Springer), editor, *Intelligent Information Systems, Bystra, Poland*, June 2000.

[5] W. Duch, T. Kucharski, J. Gomuła, and R. Adamczak. *Metody uczenia maszynowego w analizie danych psychometrycznych. Zastosowanie do wielowymiarowego kwestionariusza osobowości MMPI–WISKAD*. Toruń, Poland, 1999.

[6] W. Duch and A. Naud. Multidimensional scaling and kohonen's self-organizing maps. In *Proceedings of the Second Conference "Neural Networks and their Applications", Szczyrk*, Vol. I, pp. 138–143, 1996.

[7] W. Duch and A. Naud. On global self-organizing maps. In *Proceedings of the 4th European Symposium on Artificial Neural Networks, Bruges*, pp. 91–96, 1996.

[8] T. Kohonen. *Self-Organizing Maps*. Springer-Verlag, Heidelberg Berlin, 1995.

[9] M. A. Kraaijveld, J. Mao, and A. K. Jain. A nonlinear projection method based on kohonen's topology preserving maps. *IEEE Transactions on Neural Networks*, 6 (1995) 548–559.

[10] J. B. Kruskal. Non metric multidimensional scaling : a numerical method. *Psychometrika*, 29 (1964) 115–129.

[11] R. C. T. Lee, J. R. Slagle, and H. Blum. A triangulation method for the sequential mapping of points from n-space to two-space. *IEEE Transactions on Computers*, 26 (1977) 288–292.

[12] D. Michie, D. J. Spiegelhalter, and C. C. Taylor, editors. *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, New York, 1994.

[13] E. Pekalska, D. de Ridder, R. P. Duin, and M. A. Kraaijveld. A new method of generalizing sammon mapping with application to algorithm speed-up. In M. Boasson, J. Karndorp, J. Torino, and M. Vosselman, editors, *ASCI'99 Proc. 5th Annual Conference of the Advanced School for Computing and Image*, 1999.

[14] J. W. Sammon. A nonlinear mapping for data analysis. *IEEE Transactions on Computers* 5 (1969) 401–409.

[15] D. F. Swayne, D. Cook, and A. Buja. Xgobi: Interactive dynamic data visualization in the x window system. *AT&T Labs* Technical Rep. 1998. http://www.research.att.com/ andreas/xgobi/.

[16] M. E. Tipping. *Topographic mappings and feed-forward neural networks*. PhD thesis, Aston University, Birmingham, UK, February 1996.

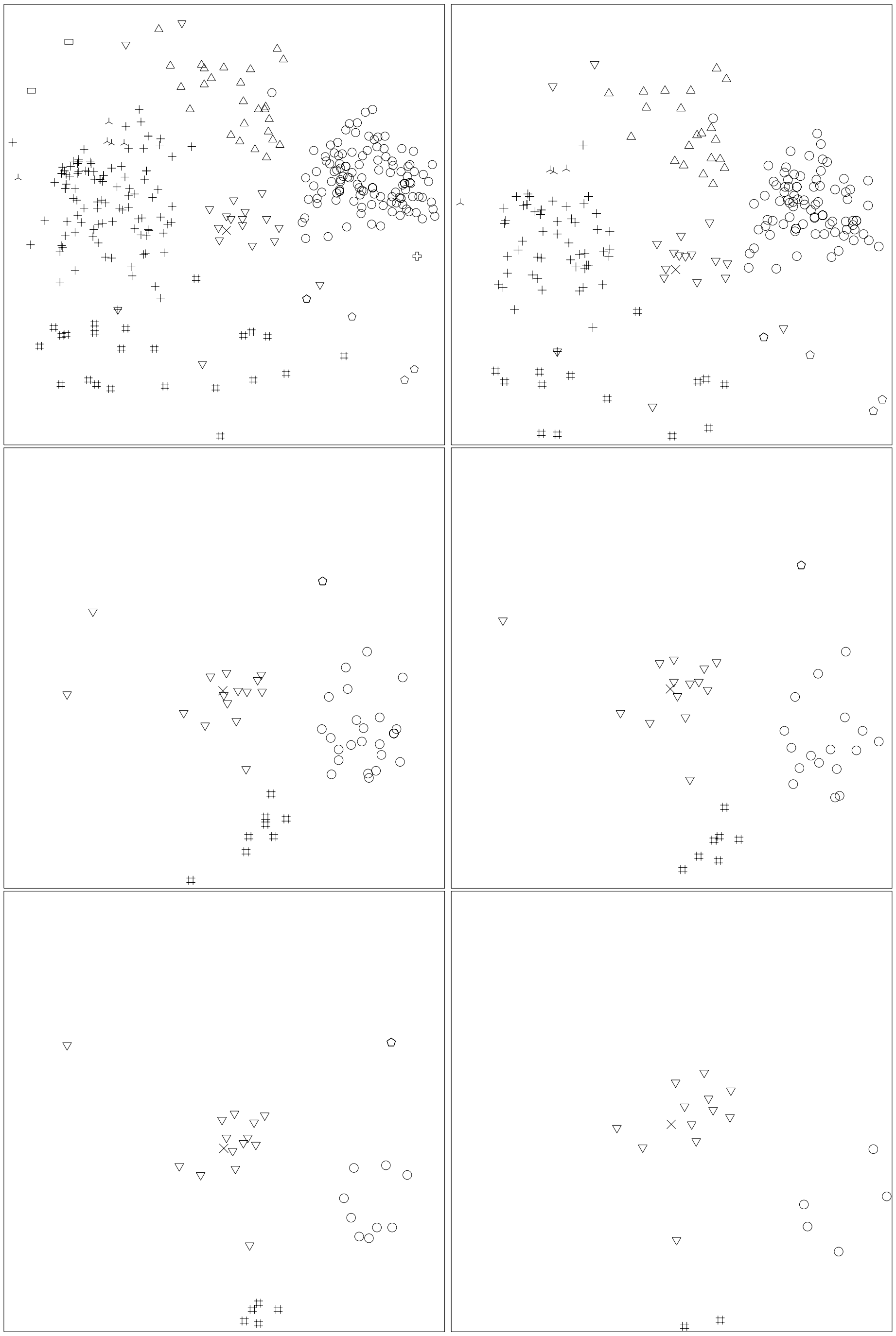

Figure 1: Zooming on the chosen case in the psychometric database. From top left to bottom right the number of points mapped (Stress values) are: 300 (0.931), 200 (0.926), 50 (0.893), 40 (0.887), 30 (0.883) and 20 (0.868).