DISTAL POINTS VIEWED IN KOHONEN'S SELF-ORGANIZING MAPS

Bartkowiak A.

Key words: Multivariate data visualization, self-organizing maps, outliers, distal points.

Abstract: Kohonen's self-organizing maps are a recognized tool for finding representative data vectors and clustering the data. To what extent is it possible to preserve the topology of the data in the constructed planar map? We address the question looking at distal data points located at the peripherals of the data cloud and their position in the provided map. Several data sets have been investigated; we present the results for two of them: the Glass data (dimension $d = 7$) and the Ionosphere data (dimension $d = 32$). It was found that the distal points are reproduced either at the edges (borders) of the map, or at the peripherals of dark regions visualized in the maps.

1 Problem

Kohonen's self-organizing maps [5, 12] are a recognized tool for finding representative data vectors and clustering the data. The method provides also a 'map' of the data. It is expected that the map will preserve a large part of the topology from the data space. To what extent is it possible to preserve the topology of the data in the constructed planar map? After all, the data are multidimensional, and the map has only two dimensions! Various neighborhood relations have been investigated, see [11, 4] and the references therein. However, little attention has been paid to the outliers and distal points of the data and how are they reproduced in the map. Maruzabál and Muñoz [6] have discussed extensively that problem and concluded that self-organizing maps are not a good tool for identifying outliers. An assessment of Kohonen's method was given by Morlini [7].

Since then, we got a graphical tool, the UMAT technique, which permits to represent distances in a planar map by smoothed color hues. The UMAT technique was already implemented in Kohonen's SOM_PAK package (1995); the technique is also available in the Matlab `SomTB2` package [13]. This gives us a powerful tool to analyze the representation of the data in the map.

The aim of the paper is to investigate, how the data points located at the peripherals of the data cloud are represented in the map. Such points are also called 'distal points', see Wouters et al., Biometrics 2003, p. 1136. The distal points might be outliers, or alternatively, they might be just extreme or peripheral points of the data cloud. Several data sets have been investigated. We present here the results for two of them: the Glass data, which contain possibly four outliers (this was not further pursued), and the Ionosphere data, which seem to be a mixture with no apparent outliers.

For each of the investigated data sets we have found the top 20 distal points; this was done using robust Mahalanobis distances [9]. Next the found 20 points were projected to the map and their position in the map was observed. It was stated that the distal points are reproduced either at the edges (borders) of the map, or at the peripherals of dark regions visualized in the map.

The paper is scheduled as follows: In section 2 we present briefly the methods. Sections 3 and 4 contain the detailed analyzes for the Glass and the Ionosphere data. Section 5 contains some concluding remarks.

2 Methods

We consider data vectors $\mathbf{x}_i,\ i = 1, \ldots, N$, with $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})$. Thus d denotes the number of variables (dimensionality), and N is the number of data vectors. The size of the data set is: $N \times d$.

Determining the distal points

The distal points were determined using robust Mahalanobis distances evaluated using the `fastmcd` algorithm developed by Rousseeuw and van Driessen [9]. For calculations we have used the Matlab function `fastmcd` offered by the cited authors. The function yields (a.o.) an index-plot of the robust Mahalanobis distances. Let $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),\ i = 1, \ldots, N$. It may be shown that squared Mahalanobis distances, evaluated for the data vectors $\mathbf{x}_i$, are distributed as χ^2_d.

Let $\chi^2_{.975}$ denote the (upper) 0.975 quantile of the χ^2_d distribution. Data points with Mahalanobis distances greater then $\sqrt{chi^2_{.975}}$ are suspected to be outliers.

Constructing the map

The methodology of Kohonen's self-organized maps id described in many sources, see, e.g., [5, 12, 6]. For the calculations we have used the Matlab `SomTB2` package [13]. Before starting the analysis, the data were standardized to mean equal zero and variance equal 1.

The basis for a self-organizing map is a lattice of given size. We have chosen a rectangular lattice composed of square units. The nodes of the grid contain neurons that are able to learn the distribution of the data. Let M denote the number of neurons. We have: $M = m_1 \times m_2$, with m_1 and m_2 denoting the side-lengths of the lattice. Each neuron designates a regular (in our case: square) map unit. The neurons are in one–to–one correspondence with prototypes of the data (called by Kohonen 'code-book vectors') located in the data space R^d.

During the process of learning, the code-book vectors adapt themselves to the distribution of the data. At the end of the learning, the entire data space

is subdivided into adjacent regions; each region contains one representative code-book vector. Each data point has assigned its nearest code-book vector.

We may display in the map various information about the data. For instance, we may show, by appropriate shadowing of the map, how distant are the prototypes in the data space. The hues depend on the chosen color-map. Dark hues may, e.g., indicate that the respective prototypes are distant, bright hues may mean that the prototypes are close.

It might be also interesting to know how many data vectors are represented by each prototype. This information may be obtained using the technique 'som_hits', shortly: 'hits'. We pass through the data and find for each data vector $\mathbf{x}_i$, $(i = 1, \ldots, N)$ its closest prototype. Say, for $\mathbf{x}_i$ this is the prototype no. h. It is in correspondence with the neuron no. h, which in turn is contained in map unit no. h. Shortly it is said that the map unit no. h was hit by the vector $\mathbf{x}_i$. After passing through the entire set of data we obtain the *counts*, an array of size [1:M], memorizing how many times a map unit was 'hit' by subsequent data vectors. The same may be done for a subset of the data (in our case: the subset of distal points), or for a different set of data with the same dimension d. The counts (number of hits) may be also displayed in the map: either in the form of properly enlarged markers, or in the form of labels expressing digitally the number of hits. We will display them in a template of the map.

3 Visualization of the Glass data

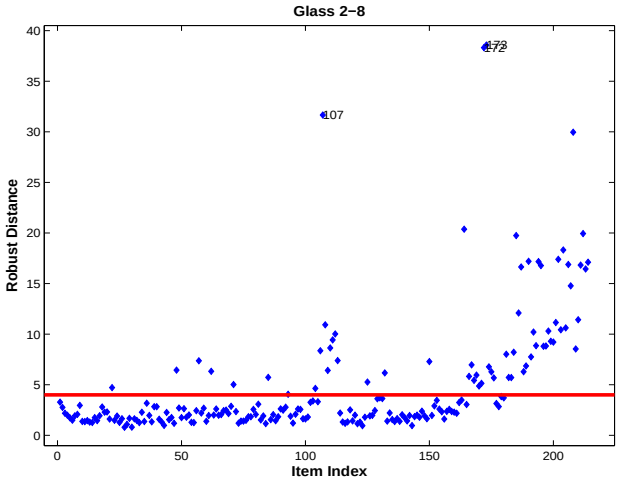

Figure 1. *Robust Mahalanobis distances for the Glass data. The top four outlying points are: no.s 173, 172, 107, 208.*

The Glass data (source: [10]) contains $N = 214$ data vectors, each characterized by $d = 7$ variables (from the original data we have taken for our

analysis the variables 2–8). The data exhibit a high multivariate kurtosis (the excess kurtosis $G_2 = 142.67$).

Index-plot for robust Mahalanobis distances calculated by the `fastmcd` procedure [9] is shown in Figure 1. One may notice that there are four outstanding points, which might be eventually considered as outliers. However, there are other data points which have quite large Mahalanobis distances. We have identified the top 20 points with the largest Mahalanobis distances – in the following these 20 points will be referred to as 'distal' points.

The map constructed for the entire glass data is shown in Figure 2. Nodes of the map represent neurons, which are in one-to-one correspondence with data prototypes located in the data space. Color shades indicate for distances of corresponding prototypes in R^d, looking in east-south directions of the map. The 20 distal points are marked by big squares squares. The topological error [13] amounts $t_e = 0.084$ (in scale [0, 1]), which is quite good.

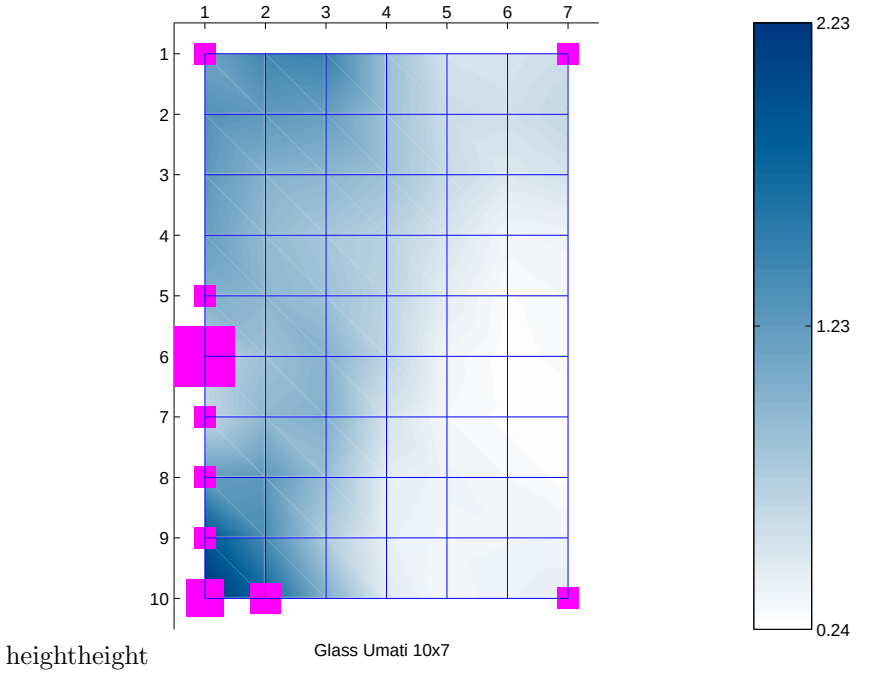

Figure 2. *Map designed by a lattice* 10×7 *obtained for the standardized Glass data. The 20 distal points are shown as big squares put over the map nodes. Notice, that all distal points appear at borders of the map.*

The *counts*, containing the number of hits received by subsequent units (see explanation in previous page) are shown in the templates below.

	Hits of 20 distal points								Remaining 194 points						
	1	2	3	4	5	6	7		1	2	3	4	5	6	7
1:	1	0	0	0	0	0	1	1:	5	1	1	3	4	2	7
2:	0	0	0	0	0	0	0	2:	2	1	0	0	5	1	2
3:	0	0	0	0	0	0	0	3:	3	2	2	1	1	2	6
4:	0	0	0	0	0	0	0	4:	3	1	3	0	5	4	6
5:	1	0	0	0	0	0	0	5:	0	1	4	0	2	3	5
6:	8	0	0	0	0	0	0	6:	0	0	1	0	5	1	2
7:	1	0	0	0	0	0	0	7:	6	1	0	3	6	3	9
8:	1	0	0	0	0	0	0	8:	5	0	1	6	0	4	10
9:	1	0	0	0	0	0	0	9:	0	1	3	3	7	1	5
10:	3	2	0	0	0	0	1	10:	0	0	1	6	7	6	3

4 Visualization of the Ionosphere data

The Ionosphere data (source: [10]) contains $N = 351$ data vectors, each characterized by 35 variables; from these the first variable is the no. of the data vector, and the last a string characterizing class membership (binary). From the original data we have taken for our analysis the variables 3–34, thus we have $d = 32$. The data exhibit a high multivariate kurtosis (the excess kurtosis amounts $G_2 = 1684.65$).

Plot of robust Mahalanobis distances for these data is shown in Figure 3. One may notice, that truly there are no outliers, only a lot of distal (peripheral) points. The composition of the data is quite interesting. It seems that we have here to deal with a mixture.

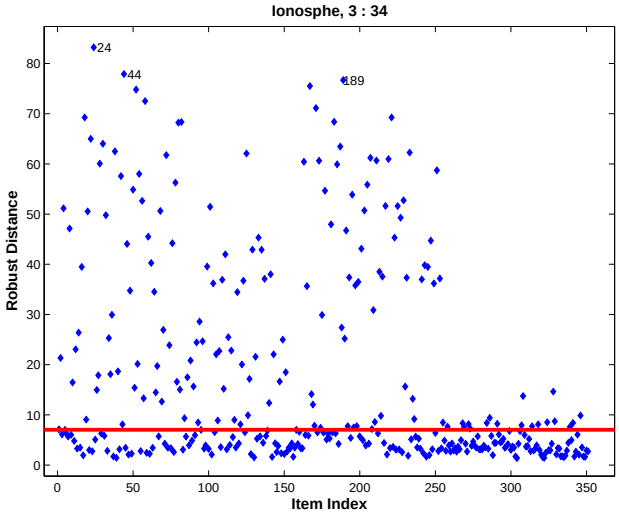

Figure 3. *Robust Mahalanobis distances for the Ionosphere data. There are no outliers, only distal points. The distribution seems to be a mixture.*

The top 20 distal points are the following ones (listed in increasing value of their Mahalanobis distances): [207 72 125 233 38 187 30 22 80 82 183 18 7221 171 58 52 167 189 44 24].

The map constructed for the entire (standardized) Ionosphere data is shown in Figure 4. The topological error [13] amounts here $t_e = 0.064$ (in scale $[0, 1]$), which is very satisfactory and indicates for a good preservation of the topology of the original data. The map units hit by the chosen 20 distal points are shown as big squares put over the nodes of the map. Distances between prototypes in R^d (when looking at east-south directions of the map) are indicated by hues of the colors, as shown in the color-bar at the right of the figure. One may notice that some of the distal points have placed themselves at the borders of the map; other distal points appear at the borders of dark regions appearing in the map.

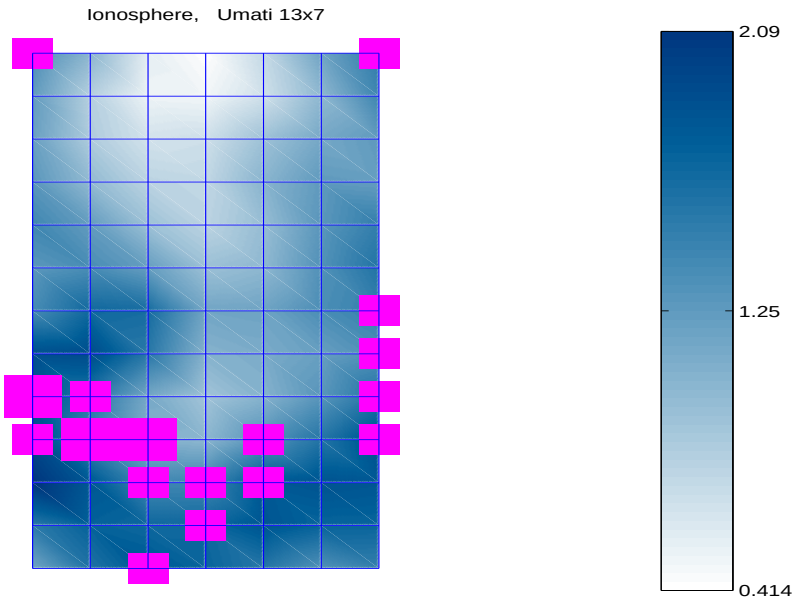

Figure 4. *SOM obtained for the standardized Ionosphere data ($d = 32$) with 20 distal points marked by big squares put over the nodes of the map. Increasing darkness indicates for increasing distance between the prototypes. Notice that distal points appear either on the borders of the map, or at the borders of dark areas.*

The *counts*, i.e. the number of hits into subsequent map units (a unit is identified by a map node) are shown in the templates below.

Top 20 distal points

	1	2	3	4	5	6	7
1:	1	0	0	0	0	0	1
2:	0	0	0	0	0	0	0
3:	0	0	0	0	0	0	0
4:	0	0	0	0	0	0	0
5:	0	0	0	0	0	0	0
6:	0	0	0	0	0	0	0
7:	0	0	0	0	0	0	1
8:	0	0	0	0	0	0	1
9:	2	1	0	0	0	0	1
10:	1	2	2	0	1	0	1
11:	0	0	1	1	1	0	0
12:	0	0	0	1	0	0	0
13:	0	0	1	0	0	0	0

71 mild distal points

	1	2	3	4	5	6	7
1:	5	0	0	0	0	1	4
2:	0	0	0	0	0	0	0
3:	1	0	0	0	0	0	0
4:	0	0	0	0	0	0	0
5:	2	0	0	0	2	0	0
6:	0	0	0	1	1	0	1
7:	1	0	0	0	1	0	3
8:	3	0	0	3	2	3	2
9:	1	1	2	1	3	3	6
10:	0	1	4	2	0	1	1
11:	1	0	1	0	4	0	0
12:	0	0	1	0	0	0	0
13:	0	0	0	1	0	1	0

Main cloud, 260 points

	1	2	3	4	5	6	7
1:	1	2	13	11	3	2	1
2:	2	1	4	5	7	2	1
3:	3	2	7	5	2	4	6
4:	3	3	3	7	5	0	7
5:	3	1	7	6	2	3	4
6:	7	3	3	5	6	1	2
7:	7	0	1	3	2	0	1
8:	0	0	0	2	0	0	0
9:	0	0	1	10	1	1	0
10:	0	0	1	0	0	1	0
11:	1	2	0	1	0	2	10
12:	7	1	3	0	0	0	2
13:	11	1	1	2	7	6	7

The entire data set was subdivided into 3 categories: 1) Top 20 distal points exhibiting the largest Mahalanobis distances in Figure 3; 2) Mild distal points with Mahalanobis distances greater then 15.0; remind that for normal data the respective 0.975 quantile equals to $\sqrt{\chi^2_{.975;32}} = 7.03$; 3) Remainder, containing 260 data points with moderate and small Mahalanobis distances (≤ 15.0).
One may notice the difference in location of hits caused by data points belonging to the 1st and 2nd categories opposed to the 3rd category.

5 Discussion and final remarks

The presented results agree with conclusions of Morlini [7]. Indeed, Kohonen's maps, when properly used, offer many highly interesting possibilities of data exploration.

Competitive methods might be: The *Neuroscale* mapping and the *Generative Topographic mapping*. The *neuroscale* mapping (see [3], also [8] and the references therein) seems to be very promising. An example of application in analysis of erosion data is shown in [2]. Preliminary comparisons of self-organizing maps (SOMs) and the GTM method are reported in [1].

References

[1] Bartkowiak A. (2003). *SOM and GTM: Comparison in Figures.* Report December 2003. `http://www.ii.uni.wroc.pl/~aba/papers.html`

[2] Bartkowiak A., Zdziarek J., Evelpidou N., Vasilopulos A. (2003). *Choosing representative data items: Kohonen, Neural Gas or Mixture Model? A case study of erosion data.* ACS'2003, The Tenth International Multiconference on Advanced Computer Systems, Szczecin-Międzyzdroje, October 22–24, 2003. Printed in disk, pp. 1 – 8. Accepted for Proceedings to be printed by Kluwer Academic Publishers.

[3] Bishop C.M., Svensèn M., and Williams C. K. I. (1996). *The generative topographic mapping.* Neural Computation **10** (1), 215 – 235.

[4] Kiviluoto K. (1996). *Topology preservation in self-organizing maps. Proceedings ICNN'96.* V. 1, IEEE Neural Networks Council, June 1996, Piscataway, New Jersey, USA, **1**, 294 – 299.

[5] Kohonen T. (1995). *Self-Organizing Maps.* Springer Series in Information Sciences, **30**, Berlin.

[6] Maruzabál, J. and Muñoz A. (1997). *On the visualization of outliers via self–organizing maps.* J. of Computational and Graphical Statistics **6**, 355 – 382.

[7] Morlini I. (1998). *Multivariate outlier detection with Kohonen's networks: an useful tool for routine exploration of large data sets.* NTSS'98, Int. Seminar on New Techniques & Technologies for Statistics, Sorrento, Italy. **2**, Contributed papers. 345 – 350.

[8] Nabney I.T. (2001). *Netlab: Algorithms for Pattern Recognition.* Springer London, Berlin, Heidelberg. Springer Series: Advances in Pattern Recognition.

[9] Rousseeuw P.J., van Driessen K. (1999). *A fast algorithm for the minimum covariance determinant.* Technometrics **41**, 212 – 223.

[10] University of California at Irvine data repository, `http://www.ics.uci.edu/pub/machine-learning-databases/`

[11] Venna J., Kaski S. (2001). *Neighborhood preservation in nonlinear projection methods: An experimental study.* In: Proceedings ICANN 2001, Vienna. Edited by G. Dorffner, H. Bischof, and K. Hornik. Springer, Berlin, 485 – 491.

[12] Vesanto J. (1999). *SOM–based data visualizing methods.* Intelligent Data Analysis, **3**(2), 111 – 126.

[13] Vesanto J., Himberg J., et al. (2000). *SOM Toolbox for Matlab 5.* SOM Toolbox Team, HUT, Finland, Libella Oy, Espoo, 1 – 54. `http://www.cis.hut.fi/projects/somtoolbox`, Version 0beta 2.0, Nov. 2001.

Address: Institute of Computer Science, University of Wrocław, Przesmyckiego 20, Wrocław 51-151, Poland

E-mail: `aba@ii.uni.wroc.pl`