ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Cartogram representation of the batch-SOM magnification factor

Alessandra Tosi[1] and Alfredo Vellido[1] *

1- Dept de Llenguatges i Sistemes Informàtics - Universitat Politècnica de Catalunya
C. Jordi Girona, 1-3, Edifici Omega, Campus Nord, 08034 Barcelona - Spain

Abstract. Model interpretability is a problem of knowledge extraction from the patterns found in raw data. One key source of knowledge is information visualization, which can help us to gain insights into a problem through graphical representations and metaphors. Nonlinear dimensionality reduction techniques can provide flexible visual insight, but the locally varying representation distortion they produce makes interpretation far from intuitive. In this paper, we define a cartogram method, based on techniques of geographic representation, that allows reintroducing this distortion, measured as a magnification factor, in the visual maps of the batch-SOM model. It does so while preserving the topological continuity of the representation.

1 Introduction

Practical data analysis methods based on machine learning and related approaches should aim to achieve a dual target, encompassing good performance (as objectively quantified, for instance, by precision, accuracy, or predictive ability) and interpretability. The latter component of this target is too often sacrificed in favour of the former.

Unfortunately, the end user of these methods in many practical application fields cannot compromise the interpretability component of the results, which, in fact, may turn out to be more important than optimum performance. This is often the case in business, industry and biomedicine, to name just a few.

Interpretability is, in the end, a problem of knowledge extraction from the patterns that can be found in raw data. Knowledge extraction may come in many flavours, one of which is information visualization. As stated in [1], information visualization can help us to gain insights into a problem through graphical metaphors, in a uniquely inductive manner that taps into the sophisticated visual capabilities of human cognition. This may be crucial when modeling complex multivariate data.

Some relevant machine learning contributions to multivariate data visualization have stemmed from the field of nonlinear dimensionality reduction (NLDR) [2]. A very popular NLDR method for visualization is the Self-Organizing Map (SOM) [3] in its many variants. This method attempts to model data through a discrete version of a low-dimensional manifold consisting of a topologically ordered grid of cluster centroids. The nonlinearity of a method such as SOM entails the existence of local distortion (magnification) in the mapping of the

*This research was funded by Spanish MICINN TIN2009-13895-C02-01 research project.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

data from the observed space into the visualization space. This involves manifold stretching and compression effects that will make the direct interpretation of the visualization a difficult undertaking.

In this study, we propose a method to reintroduce the local magnification into the low-dimensional data visualization provided by the batch-SOM algorithm. By reintroducing this distortion explicitly, we obtain a visualization that reflects the observed data more faithfully. For that, we draw inspiration from a technique originally devised for the analysis of geographic information, namely density-equalizing maps, or cartograms [4]. Cartograms are geographic maps in which the sizes of regions appear distorted in proportion to underlying quantities such as their population. Here, we replace geographical maps by batch-SOM maps and distort them according to the magnification factors (MF), which can be explicitly calculated for this model [5]. Cartograms are ideally suited to the discrete manifold representation provided by this technique.

2 Methods

2.1 Cartograms

Cartograms are cartography maps in which specific areas, often delimited by political borders, are locally distorted (stretched or compressed) to account for locally-varying underlying quantities of interest, such as population density or socio-economic data. This geometrical distortion takes (in 2-D) the form of a continuous transformation from an original plane to a transformed one, so that a vector $\mathbf{x} = (x^1, x^2)$ in the former is mapped onto the latter according to $\mathbf{x} \rightarrow T(\mathbf{x})$, in such a way that the Jacobian of the transformation is proportional to an underlying *distorting variable* $\mathbf{d}$:

$$\frac{\partial\left(T_{x^1}, T_{x^2}\right)}{\partial\left(x^1, x^2\right)} \propto \mathbf{d}. \tag{1}$$

A computationally-feasible approach to this map distortion process requires discretizing the plane continuum (and the corresponding distorting variable) to conform a regular grid. The distorting variable is assumed to take a uniform value over each of the plane fragments defined by the grid.

A method for cartogram building based on the physics principle of linear diffusion processes was recently proposed in [4]. In this method, the distorting variable $\mathbf{d}$ is let to diffuse over the map *over time* so that the final result, for $t \rightarrow \infty$, is a map of uniform distortion in which the original locations have displaced according to the process (if $\mathbf{d}$ is population density, the resulting maps are density-equalizing cartograms). The diffusion equation can be generalized to consider distortion diffusion in the form

$$\nabla^2 \mathbf{d} - \frac{\partial \mathbf{d}}{\partial t} = 0, \tag{2}$$

which has to be solved for $\mathbf{d}(\mathbf{x}, t)$, assuming that the initial condition corresponds to each map fragment being assigned its value of $\mathbf{d}$. The distortion diffusion

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

velocity can be calculated as $\mathbf{v}(\mathbf{x},t) = -\frac{\nabla \mathbf{d}}{\mathbf{d}}$ and, from it, the map location displacement as a result of which the cartogram is generated:

$$\triangle \mathbf{x} = \int_0^t \mathbf{v}(\mathbf{x},t')dt'. \tag{3}$$

To avoid arbitrary diffusion through the overall map boundaries, the map is assumed to be surrounded by an area in which the distortion is set to be the mean distortion of the complete map. This guarantees that the total map area remains constant.

2.2 The batch-SOM and its cartogram representation

A SOM consists of a layer (map) of units (or neurons) arranged in a low dimensional regular grid (often 2D, as in the current study). Each of these neurons k ($k = 1,\ldots,K$) is assigned a d-dimensional reference vector $\mathbf{y}_k$. Summarily, the algorithm proceeds by finding, for each input data point $\mathbf{x}_j$ ($j = 1,\ldots,N$) the best matching unit (BMU) $\mathbf{y}_{k_j}$ of index k_j computed as $k_j = argmin_k \{d(\mathbf{x}_j,\mathbf{y}_k)\}$. The distance $d(\cdot,\cdot)$ is often chosen to be the Euclidean one $d(\mathbf{x}_j,\mathbf{y}_k) = \|\mathbf{x}_j - \mathbf{y}_k\|$. The locations of the reference vectors are iteratively updated according the rule:

$$\mathbf{y}_k^{(t+1)} = \mathbf{y}_k^{(t)} + \alpha^{(t)} h^{(t)}(\mathbf{u}_k,\mathbf{u}_{k_j}) \left(\mathbf{x}_j^{(t)} - \mathbf{y}_{k_j}^{(t)}\right)$$

where $\mathbf{x}_j^{(t)} \in X$ is randomly selected at iteration t, $\mathbf{u}$ is the location of a unit in the SOM map, and $0 \leq \alpha^{(t)} \leq 1$ is a learning rate. The neighbourhood function $h(\cdot,\cdot)$ controls the way reference vectors other than the BMU are updated . Different alternatives can be considered, such as Gaussian h_g or step h_s functions:

$$h_g(\mathbf{u}_k,\mathbf{u}_{k_j}) = e^{-\frac{d^2\left(\mathbf{u}_k,\mathbf{u}_{k_j}\right)}{2\sigma^2}} \qquad h_s(\mathbf{u}_k,\mathbf{u}_{k_j}) = \begin{cases} 0 & if\ d(\mathbf{u}_k,\mathbf{u}_{k_j}) > \lambda \\ 1 & if\ d(\mathbf{u}_k,\mathbf{u}_{k_j}) \leq \lambda \end{cases}$$

We choose to use Gaussian neighbourhood function.

The standard SOM algorithm updates the model parameters for each data point, whereas its batch version, the batch-SOM, makes the update with the complete data set. The reference vector is now updated according to:

$$\mathbf{y}_k^{(t+1)} = \sum_{j=1}^{N} \frac{h^{(t)}(\mathbf{u}_k,\mathbf{u}_{k_j})}{\sum_{j'=1}^{N} h^{(t)}(\mathbf{u}_k,\mathbf{u}_{k_{j'}})} \mathbf{x}_j$$

where $\mathbf{u}_{k_j}$ is the node corresponding to the BMU for $\mathbf{x}_j$. This update equation can be rewritten in a kernel regression form [6], for a given iteration, as:

$$\mathbf{y}_k = \sum_{k'} \left(F(\mathbf{u}_k,\mathbf{u}_{k'})\bar{\mathbf{x}}_{k'}\right)$$

where $\bar{\mathbf{x}}_{k'} = \frac{1}{N_{k'}} \sum_{j \in G'_k} \mathbf{x}_j$ is the mean of the group $G_{k'}$ of $N_{k'}$ data points assigned to a given node k', and

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

$$F(\mathbf{u}, \mathbf{u}_k) = \frac{N_k h(\mathbf{u}, \mathbf{u}_k)}{\sum_{k'} N_{k'} h(\mathbf{u}, \mathbf{u}_{k'})}$$

2.2.1 Magnification factors for the batch-SOM

Despite the fact SOM is a discrete DR technique, in the sense that only a finite number of map units is considered, the Gaussian neighbourhood function is continuous and differentiable over the representation map space, so that calculating the local distortion (MF) over the continuum of the representation map is possible.

As described in [5], the MF represents the nonlinear distortion generated by the projection of the observed data onto the representation map. From a geometrical point of view, it can be quantified as the ratio between the area of an infinitesimal rectangle *dA'* in the representation map and the corresponding area *dA* in the observed data space

$$dA'/dA = \sqrt{det(JJ^T)},$$

where J is the Jacobian $(2 \times d)$ of the mapping transformation, whose rows are:

$$\frac{\partial \mathbf{y}}{\partial u^i} = \sum_k \left(\frac{\partial F(\mathbf{u}, \mathbf{u}_k)}{\partial u^i} \bar{\mathbf{x}}_k \right) = \sum_k \frac{u^i - u^i_k}{\sigma^2} \left(F(\mathbf{u}, \mathbf{u}_k)^2 - F(\mathbf{u}, \mathbf{u}_k) \right) \bar{\mathbf{x}}_k$$

2.2.2 Cartogram representation of the batch-SOM MF

The visualization of the MF on the batch-SOM map may inform us of the existence of data clusters and the sparsely populated spaces that separate them, as they undergo different levels of distortion: low in dense areas, while high in empty ones. In this task, it is a principled alternative to the widely used U-Matrix [7]. This direct visualization is not always intuitive. Instead, the cartogram-based representation of the batch-SOM map retains its simplicity while visually factoring out the nonlinear distortion as measured by the MF.

In the following experiments, the batch-SOM maps are transformed into a cartogram by using the rectangular grid, defined by the squares centered on the nodes, as map internal boundaries (effectively, defining a centroidal Voronoi tesselation [8]) and assuming that the level of distortion in the space beyond this rectangle is uniform and equal to the mean distortion over the complete map. Extensions to alternative grid configurations should be straightforward.

3 Experiments and discussion

We illustrate the cartogram representation of the MF for the batch-SOM with an experiment using artificial data. A total of 1,500 3-D points were randomly drawn from 3 spherical Gaussians (500 points each), all with unit variance, and with centres sitting at the vertices of a triangle. 3-D data were chosen to allow the direct visualization of the reference vectors in the observed data space.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Batch-SOM was implemented in Matlab®, using the SOM-Toolbox[1]. We used a 20×20 lattice and a Gaussian neighbourhood function. The MF and the U-matrix were calculated and cartograms were generated using these values and the lattice. For the U-matrix we used, for each unit, the average of the distances to neighbouring units.

All results are displayed in Fig.1. They include the standard batch-SOM map in the top row, left. It reflects a neat but narrow separation between the three clusters. In fact, they are far from each other, as evidenced by the direct data visualization (top row, right). The overlaid grid of reference vectors (which will not be available for data of higher dimensionality) explains this effect: many reference vectors are squashed in data-dense regions whereas only a few are stretched over the empty space in-between. This varying distortion is nicely reflected by both the MF (center, left) and the U-matrix (bottom, left).

The batch-SOM map and the distortion measures finally come together in the cartograms (MF: center, right, and U-matrix: bottom, right). The empty spaces between clusters are now fairly stretched, providing a clear view of the separation. Interestingly, part of the data reside in stretched areas: These are the ones further from the cluster centers. This effect should warn us against a too straightforward interpretation of high-distortion areas as empty spaces.

In conclusion, we have proposed and preliminary assessed a method of cartogram representation of mapping distortion for batch-SOM data visualizations. An advantage of this method is its *portability*, as it should be easy to implement for different representation architectures and with alternative NLDR visualization techniques for which distortion can be quantified.

References

[1] A. Vellido, J.D. Martín, F. Rossi and P.J.G. Lisboa, Seeing is believing: The importance of visualization in real-world machine learning applications. In M. Verleysen, editor, *proceedings of the* ESANN 2011, d-side pub., pages 219-226, Bruges (Belgium), 2011.

[2] J.A. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction*, Information Science and Statistics, Springer, 2007.

[3] T. Kohonen. *Self-Organizing Maps*, (3^{rd} ed.) Information Science Series, Springer, 2000.

[4] M.T. Gastner and M.E.J. Newman, Diffusion-based method for producing density-equalizing maps, *Proceedings of the National Academy of Sciences of the United States of America*, 101(20):7499-7504, National Academy of Sciences, 2004.

[5] C.M. Bishop, M. Svensén and C.K.I. Williams, Magnification factors for the SOM and GTM algorithms. In Proceedings of the Workshop on Self-Organizing Maps (WSOM'97), pages 333-338, June 4-6, Helsinki (Finland), 1997.

[6] F. Mulier and V. Cherkassky, Self-organization as an iterative kernel smoothing process. *Neural Computation*, 7(6):1165-1177, MIT Press, 1995.

[7] A. Ultsch, U*-Matrix: A tool to visualize clusters in high dimensional data, Technical Report, University of Marburg, 2003.

[8] Q. Du, V. Faber and M. Gunzburger, Centroidal Voronoi tessellations: Applications and algorithms, *SIAM Review*, 41(4):637-676, SIAM, 1999.

[1] www.cis.hut.fi/somtoolbox

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

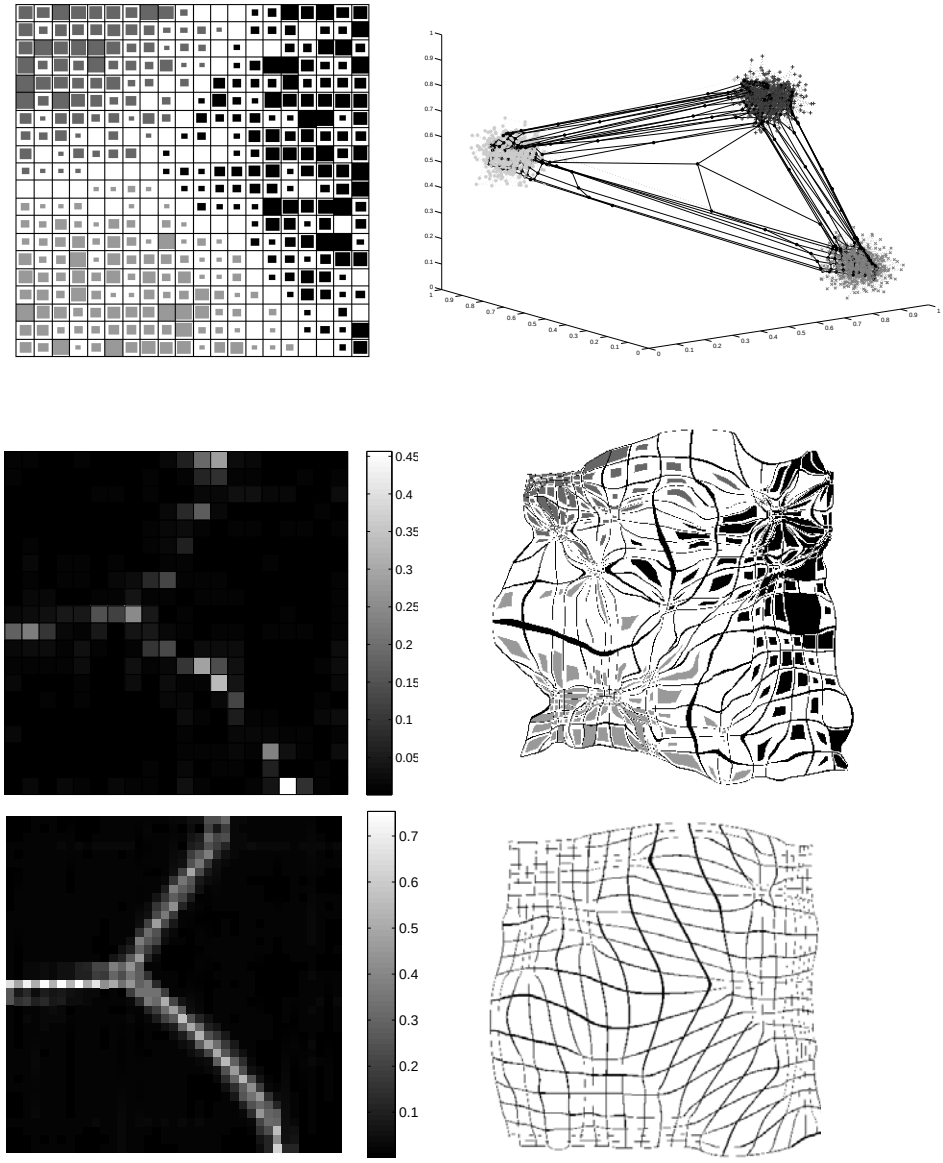

Fig. 1: Top row: left) Batch-SOM map with the BMUs of the 1,500 data points; the sizes of the squares are proportional to the number of data points assigned; different clusters are displayed in different shades of gray; right) Data points (different symbols for different clusters) overlaid with grid of reference vectors (as black dots). Center row: left) Map of MF values together with a colorbar on the righthand side of the map; right) corresponding cartogram. Bottom row: left) U-matrix map; right) corresponding cartogram.