ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

Robust cartogram visualization of outliers in manifold learning

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

1- Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, Edifici Omega, Campus Nord, 08034, Barcelona - Spain

Abstract. Most real data sets contain atypical observations, often referred to as outliers. Their presence may have a negative impact in data modeling using machine learning. This is particularly the case in data density estimation approaches. Manifold learning techniques provide low-dimensional data representations, often oriented towards visualization. The visualization provided by density estimation manifold learning methods can be compromised by the presence of outliers. Recently, a cartogram-based representation of model-generated distortion was presented for nonlinear dimensionality reduction. Here, we investigate the impact of outliers on this visualization when using manifold learning techniques that behave robustly in their presence.

1 Introduction

Most multivariate data sets generated by real application problems contain atypical observations, often referred to as outliers. What constitutes an atypical case is, indeed, a problem-dependent decision but, in any case, the presence of outlier cases is likely to have a negative impact in data modeling using machine learning and computational intelligence methods [1]. This should be particularly the case in data density estimation approaches.

Manifold learning techniques for nonlinear dimensionality reduction provide low-dimensional data representations, often with the aim of providing exploratory visualization for high-dimensional data. By forcing the model to account for their presence, outliers can compromise the results of density estimation manifold learning methods. This is likely to have an effect on the visualization they can provide.

Nonlinear manifold learning methods generate a varying local distortion that can turn exploratory visualization into a difficult undertaking. Recently, a cartogram-based representation of such model-generated distortion, inspired by geographic representation, was defined for nonlinear manifold learning [2]. The cartogram-based method was illustrated using Generative Topographic Mapping (GTM, [3]), a technique for which the mapping distortion can be quantitatively estimated in the form of magnification factors (MF, [4]).

Unfortunately, the standard GTM, as a constrained mixture of Gaussians, is prone to provide poor visualization in the presence of outliers. Alternatively, a variant of GTM defined as a mixture of Student t-distributions, namely the t-GTM, has been shown to minimize the negative effect of the presence of outliers

*This research was partially funded by Spanish research project TIN2012-31377.

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

in the modelling process [5, 6]. In this brief paper, we first define and calculate the MF for t-GTM and we then investigate the impact of outliers on the cartogram-based visualization, comparing the results yielded by the standard GTM with those of t-GTM.

2 Methods

2.1 Robust topographic mapping and its magnification factors

The GTM [3] is a nonlinear manifold learning model for multivariate data exploratory visualization. It can be seen as a low-dimensional manifold-constrained mixture of distributions model in which the centres of the distributions are defined as data centroids or prototypes $\mathbf{y}_k$, which are the images of a sample of $K, k = 1, \ldots, K$ regularly spaced latent points, as mapped from the latent to the observed data space according to the mapping function $\mathbf{y}_k = \Phi(\mathbf{u}_k)\mathbf{W}$. Here, in the standard model, Φ is a set of M Gaussian basis functions ϕ_m, while $\mathbf{W}$ is a matrix of adaptive weights, estimated as part of the model training process.

The this way defined prototypes $\mathbf{y}_k$ describe a smooth manifold that envelops the observed $D-$dimensional data $X = \{\mathbf{x}_n\}_{n=1}^N$. In the presence of outliers, some of the prototypes of this constrained mixture of Gaussians will attempt to model these atypical data, thus biasing the structure of the resulting manifold. Recently, it was proposed to redefine the GTM as a mixture of Student t-distributions, the t-GTM, so as to avoid this undesired effect [5, 6].

The conditional distribution of the observed data variables, given the latent variables, $p(\mathbf{x}|\mathbf{u})$ takes the following form for t-GTM:

$$p(\mathbf{x}|\mathbf{u}, \mathbf{W}, \beta, \nu) = \frac{\Gamma(\frac{\nu+D}{2})\beta^{\frac{D}{2}}}{\Gamma(\frac{\nu}{2})(\nu\pi)^{D/2}}(1 + \frac{\beta}{\nu}\|\mathbf{x} - \mathbf{y}(\mathbf{u})\|^2)^{-\frac{\nu+D}{2}}, \quad (1)$$

where $\Gamma(\cdot)$ is the gamma function and the parameter ν (that takes value equal to 2 in our experiments) can be viewed as a tuner that adapts the divergence from normality, and β is the inverse variance of the t-GTM noise model. The latent variables can be integrated out of the conditional distribution to obtain the likelihood of the model, so that its parameters can be optimized through maximum likelihood. For details of this procedure, see [5]. From the maximum likelihood optimization, a closed expression for responsibility r_{kn} of each latent point k for the generation of observation n, or $p(\mathbf{u}_k|\mathbf{x}_n)$, is obtained. It can be used to visualize observations in the form of a "soft-mapping" *posterior mean projection* $\mathbf{u}_n^{mean} = \sum_{k=1}^K r_{kn}\mathbf{u}_k$, in such a way that a data point is mapped in the latent space according to a responsibility-weighted combination of all latent point locations, instead of a "hard-mapping" as in winner-takes-all algorithms.

The t-GTM generates a varying local distortion that can make exploratory data visualization difficult. This distortion can be quantified over the latent space continuum with MF. The relationship between a differential area dA (for a 2-D visualization) in latent space and the corresponding area element in the GTM-generated manifold, dA', can be expressed in terms of the derivatives of the basis functions ϕ_m as $dA/dA' = \det{}^{\frac{1}{2}}(\Psi\mathbf{W}\mathbf{W}^T\Psi^T)$, where $J = \Psi\mathbf{W}$ is the

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

Jacobian of the mapping transformation, u^i is the i^{th} coordinate ($i = 1, 2$) of a latent point and Ψ is a $M \times 2$ matrix with elements φ_{mi}, defined as:

$$\frac{\partial \phi_m}{\partial u^i} = \frac{\Gamma(\frac{\nu+D}{2})(-\nu - D)\beta^{\frac{D+2}{2}}}{\Gamma(\frac{\nu}{2})\pi^{D/2}\nu^{\frac{D+2}{2}}} \left(u^i - \mu_m^i\right) \left(1 + \frac{\beta}{\nu}\|\mathbf{u} - \mu_m\|^2\right)^{-\frac{\nu+D-2}{2}} \quad (2)$$

where μ_m, $m = 1, \ldots, M$ are the centres of the Student t-distributions.

2.2 Cartogram representation for *t*-GTM

A cartogram is a depiction of an internally partitioned cartography map, in which the true surfaces of the internal partitions are distorted to reflect locally-varying quantities such as population density. This distortion is a continuous transformation from an original surface to the cartogram, 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* $\mathbf{d}$. A method for the creation of cartograms, based on linear diffusion processes was defined in [7]. This method was recently adapted to the visual representation provided by nonlinear dimensionality reduction methods in [2]. In GTM, it entails replacing geographic maps by the latent visualization map, which is transformed into a cartogram using the square regular grid formed by the lattice of latent points $\mathbf{u}_k$ as map internal boundaries. It also entails replacing distorting variables such as population density by explicit distortion measures such as MF. This method was extended to the Batch-SOM model in [8]. The reader is referred to [2] for further details.

3 Experiments

The following preliminary experiments compare the effect of outliers on the cartogram representations of the MF for the standard GTM and for *t*-GTM. An artificial dataset of 3-D points was used to make the direct visualization of the reference vectors $\mathbf{y}_k$ in the observed data space possible. A total of 1,500 3-D points were randomly drawn from 3 spherical Gaussian distributions (500 points each), all with unit variance and with centres set at the vertices of an equilateral triangle. Two different subsets of outliers were added to this dataset: *A-type*) three outliers located on the normal to the imaginary plane defined by the cluster triangle that passes through its baricenter; *B-type*) three outliers located on the normal to one vertex of the imaginary triangle.

A 15×15 regular grid and the same initialization were employed both for GTM and *t*-GTM. The MF was calculated for both methods and cartograms were generated using these values. In all cartograms, it was assumed that the level of distortion in the space beyond the grid is uniform and equal to the mean distortion over the complete map: $1/K \sum_{k=1}^{K} J(\mathbf{u}_k)$, where J is the Jacobian of the transformation. Likewise, we assumed that the level of distortion within each of the squares associated to $\mathbf{u}_k$ is itself uniform. The first experiment,

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

displayed in Fig.1, corresponds to the inclusion of *A-type* outliers, while the second, displayed in Fig.2, corresponds to the inclusion of *B-type* outliers.

Despite the fact that most GTM prototypes concentrate in the three clusters, it is clear from the image in Fig.1 (top row, left) that, in the case of standard GTM, the *A-type* outliers force the manifold towards them in an unduly way. This causes a distortion that is more controlled by the outliers than by the empty space between clusters. Even though, the cartogram visualizations generated by GTM and t-GTM are rather similar. The reason for this is the artificial symmetry of the outliers location. The maps in Fig.2, corresponding to the second experiment with added $B-type$ outliers, tell a very different story. Now, the symmetry is lost and the MF of the standard GTM reflects the fact that the model stretches one of the sides of the manifold in its attempt to cover the outliers (top row, left). As a result, an artifactual high distortion appears in the top-righ corner of the MF representation map (where outliers are seen to be mapped) and biases the cartogram representation. The t-GTM, instead, ignores the outliers and respects the symmetry of the representation while restricting the manifold to the imaginary triangle defined by the three clusters. This is clearly reflected in the corresponding cartogram.

Notice that, in both experiments, the extra MF distortion introduced by the outliers makes the data representation of the data of all clusters far more compact for GTM than for t-GTM. In any case, this simple preliminary experiments illustrate how modelling methods that behave robustly in the presence of outliers are more likely to produce more faithful representations of the nonlinear mapping distortion and, as a result, more faithful data visualizations.

References

[1] A. Vellido, E. Romero, F.F. González-Navarro, Ll. Belanche-Muñoz, M. Julià-Sapé, C. Arús, Outlier exploration and diagnostic classification of a multi-centre ^{1}H-MRS brain tumour database. *Neurocomputing*, 72(13-15):3085-3097, Elsevier, 2009.

[2] A. Vellido, D. García, À. Nebot, Cartogram visualization for nonlinear manifold learning models. *Data Mining and Knowledge Discovery*, doi: 10.1007/s10618-012-0294-6

[3] C.M. Bishop, M. Svensén, C.K.I. Williams, GTM: The Generative Topographic Mapping. *Neural Computation*, 10(1):215–234, MIT Press, 1998.

[4] C.M. Bishop, M. Svensén, 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.

[5] A. Vellido, P.J.G. Lisboa, Handling outliers in brain tumour MRS data analysis through robust topographic mapping, *Computers in Biology and Medicine*, 36(10):1049-1063, Elsevier, 2006.

[6] A. Vellido, Missing data imputation through GTM as a mixture of t-distributions. *Neural Networks*, 19(10):1624-1635, Elsevier, 2006.

[7] M.T. Gastner, 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.

[8] A. Tosi, A. Vellido, Cartogram representation of the batch-SOM magnification factor. In M. Verleysen, editor,proceedings of the European Symposium on Artificial Neural Networks Computational Intelligence and Machine Learning (ESANN 2012), d-side pub., pages 203-208, Bruges (Belgium), 2012.

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

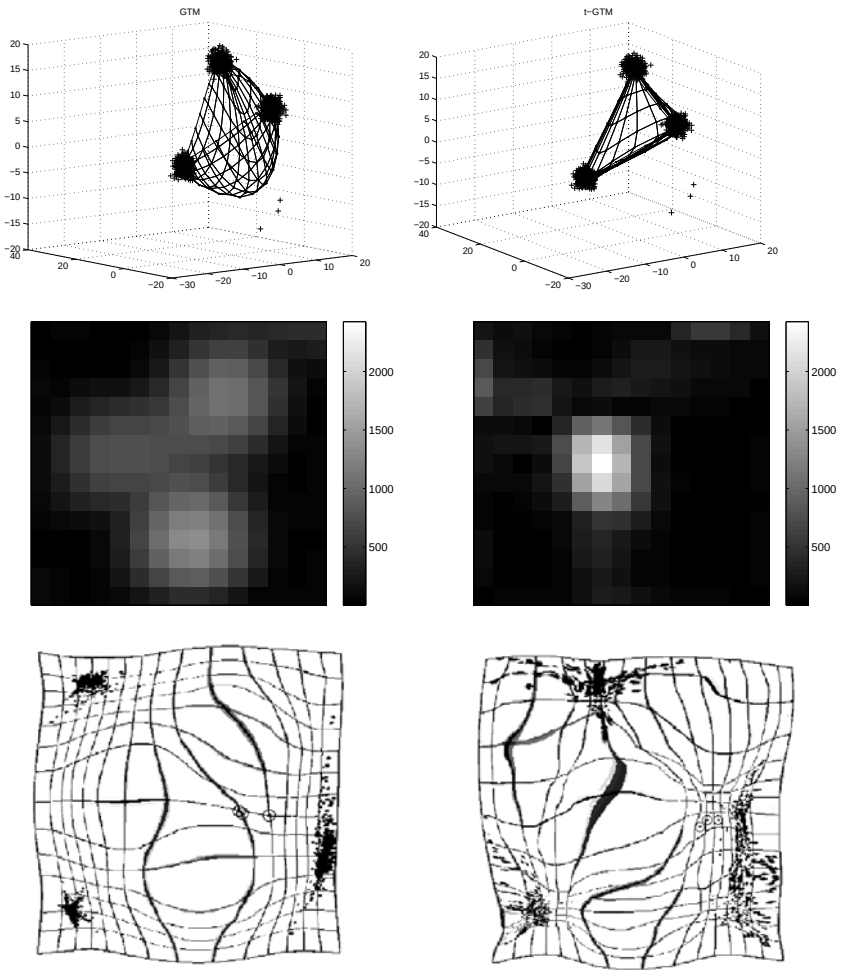

Fig. 1: Top row) Representation of the original data clusters (1,500 points, 500 in each cluster, plus $A-type$ outliers, all represented with crosses) with standard GTM (left) and t-GTM (right) The generated manifold is superimposed; it is represented as a grid whose knots are the model prototypes $\mathbf{y}_k$; central row) Maps of MF values together with a colorbar for interpretation on the right-hand side of the maps; bottom row) Corresponding cartograms, based on the MF, to which the mean projections of the data are superimposed. The mapping locations of outliers are highlighted with circles.

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

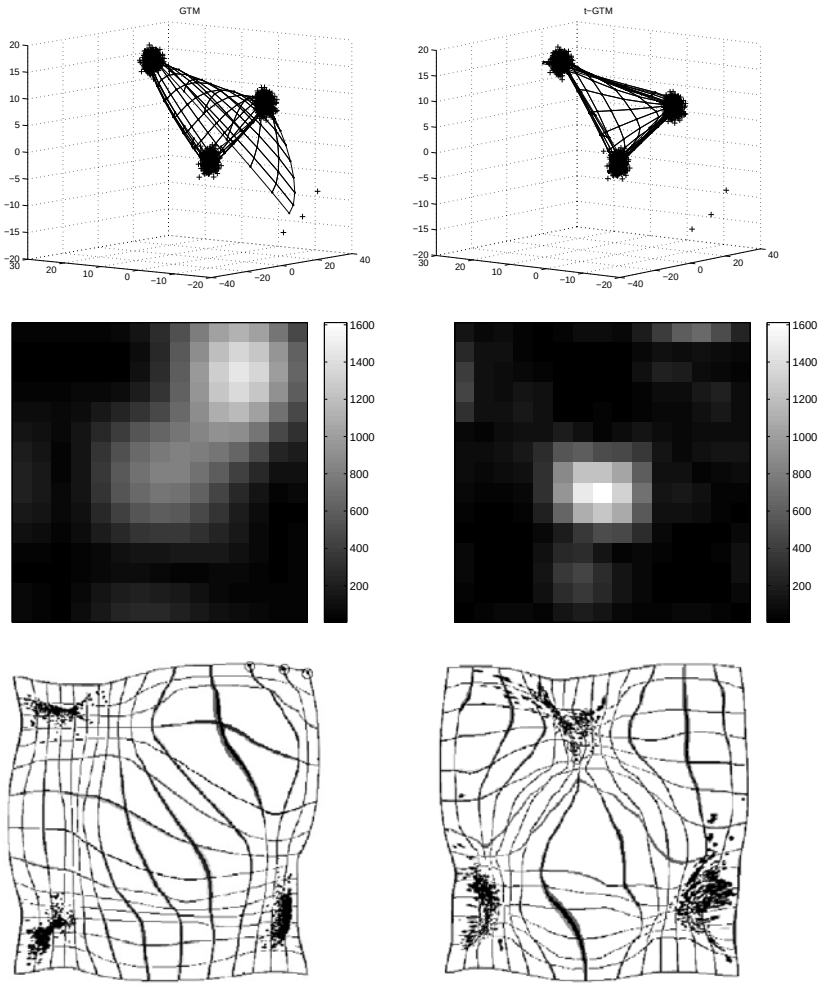

Fig. 2: Representation of data, manifold grid, MF maps and cartograms for the second experiment with $B - type$ outliers as in Fig 1. Notice the difference in the mapping locations of outliers (again highlighted with circles) as compared to Fig 1. In this case, GTM maps the outliers in a high-distortion area that is generated by the own outliers and not by the cluster data points (note that this high distortion appears because of just three outlier points), whereas the t-GTM maps them correctly to the closest cluster, without any artifactual distortion.