ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning. Bruges (Belgium), 28-30 April 2010, d-side publi., ISBN 2-930307-10-2.

Relational Generative Topographic Mapping

Andrej Gisbrecht, Bassam Mokbel and Barbara Hammer

Clausthal University of Technology - Department of Computer Science
Clausthal-Zellerfeld, Germany

Abstract. The generative topographic mapping (GTM) has been proposed as a statistical model to represent high dimensional data by means of a sparse lattice of points in latent space, such that visualization, compression, and data inspection become possible. Original GTM is restricted to Euclidean data points in a vector space. Often, data are not explicitly embedded in a Euclidean vector space, rather pairwise dissimilarities of data can be computed, i.e. the relations between data points are given rather than the data vectors themselves. We propose a method which extends the GTM to relational data and which allows to achieve a sparse representation of data characterized by pairwise dissimilarities, in latent space. The method, relational GTM, is demonstrated on several benchmarks.

1 Introduction

More and more electronic data become available in virtually all areas of life including, for example, biomedical domains, robotics, the web, or multimedia applications, such that powerful data mining tools are needed to support humans to inspect and interpret this information. Also, rapidly increasing technology such as improved sensor technology and advanced methods of data preprocessing and data storage make the data more and more complex, concerning data dimensionality and information content contained in the representation. Therefore, often, a simple comparison of data in terms of the Euclidean norm and a standard representation by means of Euclidean vectors is no longer appropriate to capture the relevant aspects of the data. Rather, dissimilarity measures which are adjusted to the data type and application area at hand should be used, including, for example, alignment distances for genomic sequence analysis in bioinformatics, the compression distance to compare texts, or structure kernels to compare complex graphs and tree structures.

Classical data mining tools such as the self-organizing map (SOM) or its statistical counterpart, the generative topographic mapping (GTM) provide a sparse representation of high-dimensional data by means of latent points arranged in a low-dimensional neighborhood structure which is useful for visualization. However, they have been introduced for Euclidean vectors only [9, 1]. Several extensions of SOM to the more general setting of data characterized by pairwise relations only, have been proposed, including median SOM which restricts prototype locations to data points [10], online and batch SOM using a kernelization of the classical approach [2, 13], and methods which rely on deterministic annealing techniques borrowed from statistical physics [5]. For GTM, a complex noise model as proposed in [12] allows the extension of the method to discrete structures such as sequences.

Recently, an intuitive extension of SOM to dissimilarity data has been proposed in [7] which relies on techniques as introduced in [8]: assume that only a dissimilarity matrix characterizes the data and an explicit vectorial representation is unknown. If prototypes have the special form of convex combinations

of data points, classical SOM can be computed indirectly by adapting the coefficient vectors without any explicit reference to the underlying vector space or a formula of the dissimilarity measure. The resulting algorithm, relational SOM, arrives at a sparse representation of dissimilarity data in terms of virtual prototypes represented by coefficient vectors.

In this contribution, we extend this principle to GTM. For this purpose, we use the trick of an indirect representation of prototypes in the image space in terms of convex combinations of data points and the associated possibility to compute distances in the space without an explicit reference to the vector representation of points. We show that an EM algorithm can be derived to obtain the parameters of the model by maximizing the data log-likelihood. The efficiency and feasibility of this method, relational GTM, is demonstrated on several benchmark data sets given by dissimilarity matrices.

2 The generative topographic mapping

The GTM [1] provides a generative stochastic model of data $\mathbf{x} \in \mathbb{R}^D$ which is induced by a mixture of Gaussians with centers induced by a regular lattice of points $\mathbf{w}$ in latent space. These are mapped to prototypical target vectors $\mathbf{w} \mapsto \mathbf{t} = y(\mathbf{w}, \mathbf{W})$ in the data space, where the function y is parameterized by $\mathbf{W}$, e.g. a generalized linear regression model $\Phi(\mathbf{w}) \cdot \mathbf{W}$ induced by base functions Φ such as equally spaced Gaussians with bandwidth σ. Every latent point induces a Gaussian distribution

$$p(\mathbf{x}|\mathbf{w}, \mathbf{W}, \beta) = \left(\frac{\beta}{2\pi}\right)^{D/2} \exp\left(-\frac{\beta}{2}\|\mathbf{x} - y(\mathbf{w}, \mathbf{W})\|^2\right) \tag{1}$$

with bandwidth β, which generates a mixture of K modes

$$p(\mathbf{x}|\mathbf{W}, \beta) = \sum_{k=1}^{K} p(\mathbf{w}_k) p(\mathbf{x}|\mathbf{w}_k, \mathbf{W}, \beta) \tag{2}$$

where $p(\mathbf{w}_k)$ is often chosen as uniform distribution. GTM training optimizes the data log-likelihood

$$\ln\left(\prod_{n=1}^{N}\left(\sum_{k=1}^{K} p(\mathbf{w}_k) p(\mathbf{x}_n|\mathbf{w}_k, \mathbf{W}, \beta)\right)\right) \tag{3}$$

with respect to $\mathbf{W}$ and β. This can be done by means of an EM approach which treats the generative mixture component $\mathbf{w}_k$ for a data point $\mathbf{x}_n$ as hidden parameter. Choosing a generalized linear regression model and uniform distribution of the latent points, EM training in turn computes the responsibilities

$$R_{kn}(\mathbf{W}, \beta) = p(\mathbf{w}_k|\mathbf{x}_n, \mathbf{W}, \beta) = \frac{p(\mathbf{x}_n|\mathbf{w}_k, \mathbf{W}, \beta) p(\mathbf{w}_k)}{\sum_{k'} p(\mathbf{x}_n|\mathbf{w}_{k'}, \mathbf{W}, \beta) p(\mathbf{w}_{k'})} \tag{4}$$

of component k for point number n, and the model parameters by means of the formulas

$$\mathbf{\Phi}^t \mathbf{G}_{\text{old}} \mathbf{\Phi} \mathbf{W}^t_{\text{new}} = \mathbf{\Phi}^t \mathbf{R}_{\text{old}} \mathbf{X} \tag{5}$$

for $\mathbf{W}$ where $\mathbf{\Phi}$ refers to the matrix of base functions, $\mathbf{X}$ to the data points, $\mathbf{R}$ to the responsibilities, and $\mathbf{G}$ is a diagonal matrix with accumulated responsibilities $G_{nn} = \sum_n R_{kn}(\mathbf{W}, \beta)$. The bandwidth can be computed by solving

$$\frac{1}{\beta_{\text{new}}} = \frac{1}{ND} \sum_{k,n} R_{kn}(\mathbf{W}_{\text{old}}, \beta_{\text{old}}) \|\mathbf{\Phi}(\mathbf{w}_k)\mathbf{W}_{\text{new}} - \mathbf{x}_n\|^2 \tag{6}$$

where D is the data dimensionality and N the number of data points.

3 Relational GTM

We assume that data $\mathbf{x}$ are given only indirectly in terms of pairwise dissimilarities $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|^2$, but the vector representation of the data is unknown. Thus, for general prototypes $\mathbf{t}$, the probability (1) cannot be computed, nor is it possible to determine prototypical targets at all, if no embedding vector space is known. In [8], the following fundamental observation is presented: assume that prototypes are restricted to convex combinations of data points, i.e.

$$\mathbf{t}_k = \sum_{n=1}^{N} \alpha_{kn}\mathbf{x}_n \text{ where } \sum_{n=1}^{N} \alpha_{kn} = 1 \tag{7}$$

Then, the prototypes $\mathbf{t}_k$ can be represented indirectly by means of the coefficient vector α_k and, further, distances of data points and prototypes can be computed as in [8]

$$\|\mathbf{x}_n - \mathbf{t}_k\|^2 = [\mathbf{D}\alpha_k]_n - \frac{1}{2} \cdot \alpha_k^t \mathbf{D}\alpha_k \tag{8}$$

where $\mathbf{D}$ refers to the matrix of pairwise dissimilarities of data points and $[\cdot]_i$ is component i of the vector. This observation has been used in [7] to derive a relational variant of SOM. We show, that the same principle allows us to generalize GTM to relational data described by a dissimilarity matrix $\mathbf{D}$. We restrict prototype vectors $\mathbf{t}_k$ to the convex hull of data points and represent those in terms of coefficient vectors α_k. Hence, we can directly treat the mapping of latent points to prototype points as mapping of the latent space to the coefficients:

$$y : \mathbf{w}_k \mapsto \alpha_k = \Phi(\mathbf{w}_k) \cdot \mathbf{W} \tag{9}$$

where Φ refers to base functions such as equally spaced Gaussians with bandwidth σ in the latent space. To apply (8), we set the restriction

$$\sum_n [\Phi(\mathbf{w}_k) \cdot \mathbf{W}]_n = 1 \tag{10}$$

This way, the likelihood function (3) can be computed based on (1) where the distance computation can be performed indirectly using (8). As for GTM, we can use an EM optimization scheme to arrive at solutions for the parameters β and $\mathbf{W}$, where, again, the mode $\mathbf{w}_k$ responsible for data point $\mathbf{x}_n$ serves as hidden parameter. An EM algorithm in turn computes the responsibilities (4) using the distance (8), and it optimizes the expectation

$$\sum_{k,n} R_{kn}(\mathbf{W}_{\text{old}}, \beta_{\text{old}}) \ln p(\mathbf{x}_n | \mathbf{w}_k, \mathbf{W}_{\text{new}}, \beta_{\text{new}}) \tag{11}$$

with respect to $\mathbf{W}$ and β under the constraint (10). Lagrange optimization with Lagrange multiplier μ_k for component $\mathbf{t}_k$ leads to the equation

$$\beta\mathbf{\Phi}^T\mathbf{G}_{\text{old}}\mathbf{\Phi}\mathbf{W}_{\text{new}} - \beta\mathbf{\Phi}^T\mathbf{R}_{\text{old}}\mathbf{X} + \mathbf{\Phi}^T\mu\mathbf{1}_{1,N} = 0 \tag{12}$$

where $\mathbf{1}_{Q,L}$ denotes a $Q \times L$ matrix with entries 1, Φ is the matrix of base functions evaluated at the data points, μ is the vector of Lagrange multipliers, and the remaining symbols refer to the corresponding matrices as introduced above. After algebraic manipulations of (12), it follows that the Lagrange multipliers vanish. Hence the model parameters can be determined in analogy to (5,6). We refer to this method as relational GTM (RGTM).

Initialization uses a MDS projection of the dissimilarities to two dimensional points $\mathbf{A}$. This induces the two primary coefficients of the unit vectors in the space of convex combinations in $\mathbb{R}^N$. The weights $\mathbf{W}$ should be initialized such that the latent grid is mapped to the two-dimensional manifold spanned by these components. Since this hyperplane is going through the origin, $\mathbf{1}/N$ should be added to the linear component of $\mathbf{W}$. Normalizing the matrix such that there are no negative coefficients, we compute $\mathbf{\Phi W} = \mathbf{X}\mathbf{A}^T/(N \max_{ij}(|[\mathbf{X}\mathbf{A}^T]_{ij}|))$, and add $\mathbf{1}/N$ afterwards.

4 Experiments

First, we test RGTM on several benchmark dissimilarity data sets as introduced in [3, 6]: cat cortex (65 data points and 4 classes), patrol data (241 points, 8 classes), voting data (435 samples, 2 classes), protein data (226 points, 5 classes), aural sonar (100 points, 11 classes); in each case, a symmetric dissimilarity matrix with zero diagonal is given representing a problem-adapted measurement of the dissimilarity of data points.

Since these data sets are labeled, it is possible to evaluate the result by the classification error obtained by posterior labeling. Thereby, posterior labeling of RGTM takes place based on the majority label of the accumulated responsibility of a latent point for data points carrying this label. We report the results of a repeated cross-validation with ten repeats, where we use 2 folds for the cat cortex data and aural sonar data and 10 folds for the other data sets to maintain comparability with the results from [6]. For cross-validation, out of sample extensions of the assignments can be computed the same way as for relational neural gas, see [6]. In all cases, we use 100 latent points and 4 base functions given by Gaussians. This global parameter setting was optimized with regard to all data sets. The initial β, which determines the bandwidth of the base functions, has only a slight effect on the algorithm, if it stays in a reasonable interval. Here, the number of base functions is chosen as small as possible to preserve the topology of the data. Changing the number of latent points generally changes only the sampling of the data but the shape of the map stays the same. So with a smaller number, the algorithm is faster and sparsity of the representation is increased; with a larger number, the algorithm is slower but more details in data relations can be discovered. The classification accuracy obtained on the test set is depicted in Tab. 1. For comparison, we report the classification accuracy of deterministic annealing (DA) and relational neural gas (RNG) as presented in [6]. Obviously, RGTM is always competitive to these two alternatives and it is even better for three of the five classification tasks. Hence

	RNG	DA	**RGTM**
cat cortex	0.698 (0.076)	0.803 (0.083)	0.863 (0.027)
proteins	0.919 (0.016)	0.907 (0.008)	0.938 (0.008)
aural sonar	0.834 (0.014)	0.856 (0.026)	0.849 (0.030)
patrol	0.665 (0.024)	0.521 (0.051)	0.714 (0.034)
voting	0.950 (0.004)	0.951 (0.005)	0.942 (0.009)

Table 1: Classification results on the data sets obtained by a repeated cross validation, the standard deviation is given in parenthesis.

RGTM offers a feasible alternative to DA and RNG, where RGTM has at least the same capacity, but it is based on an explicit statistical model and displays rapid convergence of the algorithm since it is based on a fast EM scheme.

To demonstrate the visualization features of RGTM, we show the topographic mapping for a dissimilarity data set of classical music similar to [11]. It is comprised of pairwise dissimilarities between 1068 sonatas from the classical period (by Beethoven, Mozart and Haydn) and the baroque era (by Scarlatti and Bach). The musical pieces were given in the MIDI file format, taken from the online MIDI collection *Kunst der Fuge*[1]. Their mutual dissimilarities were measured with the normalized compression distance (NCD), see [4], using a specific preprocessing, which provides meaningful invariances for music information retrieval. This method uses a graph-based representation of the musical pieces to construct reasonable strings as input for the NCD, see [11]. On this data, the RGTM was trained using 400 latent points and 4 base functions. Since there is no ground truth for this kind of musical dissimilarity measure, no precise interpretation and evaluation of the results is possible. Still, the visualization features of RGTM can be demonstrated in comparison to the existing RSOM, as seen in Fig. 1.

5 Conclusions

In this contribution, the generative topographic mapping has been extended towards data given by a dissimilarity matrix rather than Euclidean vectors. The resulting algorithm, relational GTM, can be used directly on the dissimilarity matrix. It has been demonstrated in the experiments that RGTM provides a reasonable topographic mapping of the data which is at least competitive if not superior to alternatives such as deterministic annealing and relational neural gas while providing an explicit stochastic model.

Note that RGTM leads to a sparse representation of data in terms of a set of latent points in latent space together with a prescription of how this generates a probability distribution in data space. Unlike standard GTM, however, the targets of latent points in the data space (the prototypes) are given only indirectly through vectors of coefficients, which are not sparse. In [6], approximation schemes have been proposed in the context of relational neural gas which, on the one hand, result in a sparse representation of prototypes, on the other hand, allow a patch processing of huge dissimilarity matrices for which the computational load would otherwise be too big. This way, the resulting topographic mapping scheme is linear in the number of data points. The transfer of this method to RGTM is subject of ongoing research.

[1] `http://www.kunstderfuge.com`

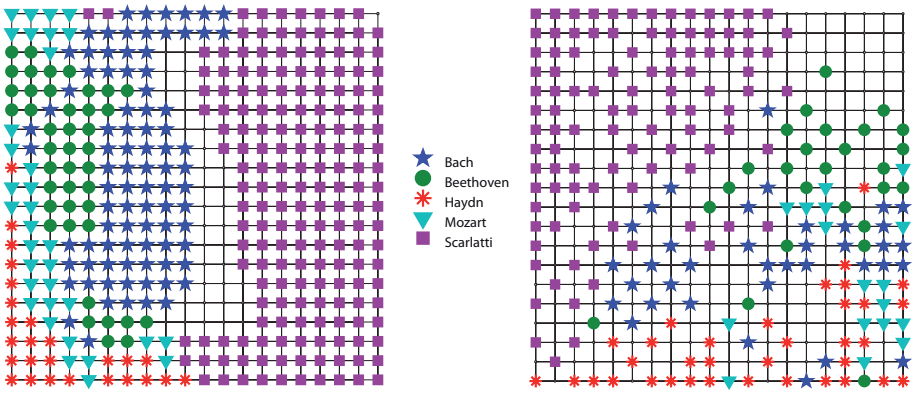

Fig. 1: RGTM (left) and RSOM (right) visualization of classical and baroque sonatas by Beethoven (102), Haydn (172), Mozart (147), Bach (92), and Scarlatti (555). The grid points are marked using posterior labeling. The RGTM grid shows a noticeable separation of the musical pieces by composer and epoch, where mostly the comprehensive work of Bach marks a blend between the classical and baroque era. The shown arrangement seems meaningful since Bach's work is considered influential for both musical eras. Also the distinct style of Scarlatti is represented. In the grid on the right, generated with RSOM (trained in 500 epochs with an initial neighborhood range of 40 on the same data set) the separation of the composers is less distinct.

References

[1] C. Bishop, M. Svensen, and C. Williams. The generative topographic mapping. *Neural Computation* 10(1):215-234, 1998.

[2] R. Boulet, B. Jouve, F. Rossi, and N. Villa. Batch kernel SOM and related Laplacian methods for social network analysis. *Neurocomputing* 71(7-9): 1257-1273, 2008.

[3] Y.Chen, E.K.Garcia, M.R.Gupta, A.Rahimi, and L.Cazzani. Similarity-based classification: concepts and algorithms, *Journal of Machine Learning Research* 10: 747-776, 2009.

[4] R.Cilibrasi and M.B.Vitanyi, Clustering by compression, *IEEE Transactions on Information Theory* 51(4):1523-1545, 2005.

[5] T. Graepel and K. Obermayer. A stochastic self-organizing map for proximity data. *Neural Computation* 11:139-155, 1999.

[6] B. Hammer and A. Hasenfuss. Topographic Mapping of Large Dissimilarity Data Sets. Technical Report, Clausthal University of Technology, IfI-10-01, January 2010.

[7] A. Hasenfuss and B. Hammer. Relational Topographic Maps. M.R. Berthold, J. Shawe-Taylor, and N. Lavrac (eds.), *IDA 2007*: 93-105, 2007.

[8] R. J. Hathaway and J. C. Bezdek. Nerf c-means: Non-Euclidean relational fuzzy clustering. *Pattern Recognition* 27(3):429-437, 1994.

[9] T. Kohonen. *Self-Organizing Maps*, Springer, 1995.

[10] T. Kohonen and P. Somervuo. How to make large self-organizing maps for nonvectorial data. *Neural Networks* 15:945-952, 2002.

[11] B. Mokbel, A. Hasenfuss, and B. Hammer. Graph-Based Representation of Symbolic Musical Data. *GbRPR 2009*, A. Torsello, F. Escolano, and L. Brun (eds.), Springer, pp. 42-51, 2009.

[12] P. Tino, A. Kaban, and Y. Sun. A generative probabilistic approach to visualizing sets of symbolic sequences. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD-2004*, R. Kohavi, J. Gehrke, W. DuMouchel, J. Ghosh (eds). pp. 701-706, ACM Press, 2004.

[13] H. Yin. On the equivalence between kernel self-organising maps and self-organising mixture density network. *Neural Networks* 19(6):780-784, 2006.