ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

Magnification Control in Relational Neural Gas

Alexander Hasenfuss[1], Barbara Hammer[1], Tina Geweniger[2], and Thomas Villmann[2]

1- Clausthal University of Technology - Department of Informatics
D-38678 Clausthal-Zellerfeld - Germany

2- University of Leipzig - Clinic for Psychotherapy
D-04107 Leipzig - Germany

Abstract. Prototype-based clustering algorithms such as the Self Organizing Map (SOM) or Neural Gas (NG) offer powerful tools for automated data inspection. The distribution of prototypes, however, does not coincide with the underlying data distribution and magnification control is necessary to obtain information theoretic optimum maps. Recently, several extensions of SOM and NG to general non-vectorial dissimilarity data have been proposed, such as Relational NG (RNG). Here, we derive a magnification control scheme for RNG based on localized learning, and we demonstrate its applicability for various data sets.

1 Introduction

SOM constitutes one of the most prominent data inspection tools with numerous applications in image processing, robotics, telecommunication, etc. [10]. Neural Gas constitutes an alternative clustering method proposed by Martinetz et al. [11], which automatically detects a data optimum lattice. Thus, clustering results are more robust, but additional steps such as MDS are required for visualization. Both algorithms, however, are developed for vectorial data and they cannot directly be applied to data such as strings, time series, graphs, etc.

Several extensions of clustering towards more general data have been proposed, see e.g. [7] for an overview. A very elegant way is to model data by pairwise (dis-)similarities. One way to deal with such data is Median clustering [1] which restricts prototype locations to given data points in standard batch optimization. However, it has the drawback that a fine-grained optimization of prototype locations is not possible. Relational SOM and NG [5, 8] extend batch clustering towards general dissimilarity data by means of the relational dual resulting in a scheme which is equivalent to standard NG and SOM for Euclidean data and which allows smooth updates in the general setting.

General vector quantization does not compute information optimum maps, rather, the prototype distribution follows the data distribution by means of a power law with magnification exponent $\neq 1$ [15]. By changing the learning rule, the magnification exponent of neural maps can be controlled [13]. This can lead to great benefits in practical applications as demonstrated in [9, 14]. Here, we extend magnification control to RNG based on a local learning scheme for Batch NG as proposed in [6]. This way, we obtain a method to control the focus of the neural map on sparse resp. dense regions of the data space for general dissimilarity data. We demonstrate the behavior for several benchmark sets.

2 Neural Gas

Neural Gas (NG) [11], is a vector quantization technique aiming for representing given data $v \in V \subseteq \mathbb{R}^d$ faithfully by prototypes $w_i \in \mathbb{R}^d$, $i = 1, \ldots, n$. For an input distribution given by a density $P(v)$, and neighborhood range controlled through the function $h_\lambda(t) = \exp(-t/\lambda)$ with $\lambda > 0$, its cost function is

ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

$$E = \frac{1}{2} \sum_{i=1}^{n} \int h_\lambda(k(w_i, v)) \cdot \|v - w_i\|^2 P(v) dv,$$

where $k(w_i, v) = |\{w_j : \|v - w_j\| < \|v - w_i\|\}|$ denotes the rank of neuron w_i arranged according to the distance from data point v. Typically, NG is optimized online. Batch optimization as proposed in [1] optimizes the cost function for a given set $\{v_1, v_2, \ldots, v_m\}$. It, in turn, determines the ranks k_{ij} for fixed w_i and new prototypes $w_i = \sum_j h_\lambda(k_{ij}) \cdot v_j / \sum_j h_\lambda(k_{ij})$ for fixed ranks k_{ij}.

3 Relational Neural Gas

We assume data are given only by means of pairwise distances d_{ij}. As already mentioned beforehand, median clustering [1] restricts prototype locations to data points such that only discrete optimization steps are possible. Relational Neural Gas [5] overcomes this problem by using convex combinations of the data points. Assume that there exists an (unknown and possibly high-dimensional) embedding of the data points in a Euclidean space, i.e. $d_{ij} = \|v_i - v_j\|$. Then, optimum prototypes can be expressed as $w_i = \sum_j \alpha_{ij} v_j$ with $\sum_j \alpha_{ij} = 1$. Therefore, quadratic distances $\|w_i - v_j\|^2$ between feature points and prototypes can be expressed as $\|w_i - v_j\|^2 = (\Delta \cdot \alpha_i)_j - 1/2 \cdot \alpha_i^t \cdot \Delta \cdot \alpha_i$, where $\Delta = (d_{ij}^2)_{ij}$ constitutes the quadratic distance matrix and $\alpha_i = (\alpha_{ij})_j$ the coefficients of prototypes. Thus, we can use this term in batch optimization to compute optimum ranks, and, in turn, we can compute optimum prototype locations given fixed ranks

$$\alpha_{ij} = h_\lambda(k_i(v_j)) / \sum_j h_\lambda(k_i(v_j)). \tag{1}$$

This allows to reformulate batch optimization in terms of relational data [5]. This scheme is equivalent to batch NG if an Euclidean embedding of the data points exists. If this is not possible, the consecutive optimization can still be applied. It has been shown in [5] that this algorithm converges for every nonsingular symmetric matrix Δ and it optimizes the relational dual cost function of NG.

4 Magnification Control

As demonstrated by Zador [15], vector quantization techniques aiming for a minimization of the distortion error feature the inherent characteristic that the final prototype density ρ does not exactly match the data density P. The relation asymptotically obeys the power law $\rho(w) \sim P(w)^\alpha$, with $\alpha = D/(D+2)$ for vector quantizers minimizing the quadratic distortion error, where D denotes the intrinsic data dimensionality. The exponent α is called *magnification exponent*.

Arbitrary magnification can be achieved, among other techniques, by a localized learning strategy [13]. The update rule is extended by a local learning rate which depends on the local data density. As derived in [6], the batch update is

$$w_i = \sum_j \left(h_\lambda(k_{ij}) \cdot P(v_j)^m \cdot v_j\right) / \sum_j \left(h_\lambda(k_{ij}) \cdot P(v_j)^m\right). \tag{2}$$

Associated with this update is a modified magnification power law $\rho(w_i) \sim P(w_i)^{\alpha'}$ where $\alpha' = (m+1) \cdot \alpha$. Parameter m allows to control the magnification as desired. The information theoretic optimum $\alpha' = 1$ is reached at $m = 2/D$.

ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

The localized learning technique can easily be transferred to RNG by integrating the factor into the prototype updates rule (2) as follows

$$\alpha_{ij} = (h_\lambda(k_i(v_j)) \cdot P(v_j)^m) \,/\, \sum_j (h_\lambda(k_i(v_j)) \cdot P(v_j)^m)\,. \tag{3}$$

If an Euclidean embedding of data points exists, this learning rule is equivalent to local learning for batch NG (2) as can be seen by inserting the rule (3) into the prototype representation $w_i = \sum_j \alpha_{ij} v_j$. Thus, the theoretical guarantees as derived in [6] hold for this case. For the non-Euclidean case, the theoretical effect of the localized learning rule is not clear a priori, we can, however, show convergence for *every* nonsingular and symmetric Δ: Consider the cost function

$$E(k_{ij}, \alpha_{ij}) = \sum_{i,j} h_\lambda(k_{ij}) P(v_j)^m \Big(\sum_l d_{jl}^2 \alpha_{il} - \frac{1}{2} \sum_{l,l'} d_{ll'}^2 \alpha_{il} \alpha_{il'} \Big) \tag{4}$$

with some function $P(v_j)$ (not necessarily a density). Assume this cost function is optimized under the condition that k_{ij} constitutes a permutation of $\{0, \ldots, n-1\}$ for all j. It is obvious, that localized RNG computes optimum values k_{ij} for fixed α_{ij}. Conversely, for fixed k_{ij}, optimum α_{ij} obey $\partial E(k_{ij}, \alpha_{ij})/\partial \alpha_{nl} = 0$, hence

$$0 = \sum_j d_{lj}^2 \Big(h_\lambda(k_{nj}) P(v_j)^m - \sum_{j'} h_\lambda(k_{nj'}) P(v_{j'})^m \cdot \alpha_{nj} \Big)$$

for all n, l, which, for nonsingular matrix Δ, yields Eqn. (3). Hence, the function $E(k_{ij}, \alpha_{ij})$ is in turn optimized for α_{ij} and k_{ij} in localized RNG. Thus, it converges after a finite number of steps to a local optimum of Eqn. (4) because only finitely many values k_{ij} exist. One can compute that, for optimum α_{ij}, the function (4) is equivalent to the extended relational dual cost function

$$E^\vee = \frac{1}{2} \cdot \sum_i \sum_{l,l'} h_\lambda(k_{il}) h_\lambda(k_{il'}) P(v_l)^m P(v_{l'})^m d_{ll'}^2 \Big/ \sum_{l''} h_\lambda(k_{il''}) P(v_{l''})^m \tag{5}$$

which measures the dissimilarities of data points assigned to the same clusters, weighted according to $P(v_l)^m$. The denominator accounts for the fact that the size of clusters per se is not important. Obviously, the control parameter m allows to control the relevance of the value $d_{ll'}^2$ of data points v_l in certain regions of the data space. Assume $P(v_l)$ measures the relative number of similar points (or local data density, if defined). Then a control parameter $m > 1$ emphasizes regions which contain a large number of pairwise similar data points, whereas $m < 1$ emphasizes regions with only few pairwise similar points.

5 Experiments

We test local learning for a Euclidean benchmark, and four non-Euclidean settings as described in [12, 4]. In the latter cases, local learning can be applied and it can be expected that 'dense' or 'sparse' regions, respectively, of the data are emphasized depending on m due to the optimized costs (5). However, the exact theoretical law of the prototype density and its information theoretic optimum is not known. Note that, if Δ stems from a metric, the concept of dimensionality can be defined for the underlying data manifold and the intrinsic data dimensionality can been estimated using a Grassberger-Procaccia analysis [3]. Similarly,

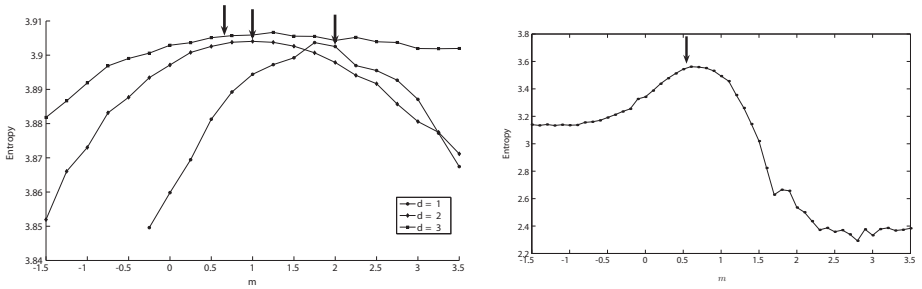

Fig. 1: *Left:* Euclidean benchmark data – Entropy of map formation for different values m of magnification control and training sets of intrinsic dimensionality $d \in \{1, 2, 3\}$ — *Right:* Chicken Pieces Dataset – Entropy for different magnification control parameter values m

density estimation is possible for separable metric spaces and uniformly continuous densities by means of histogram estimators [2]. For simplicity, we compute $P(v)$ using a simple Parzen window with bandwidth chosen as a third of the average point distance, which gives a rough approximation to the underlying density for Euclidean and metric settings, respectively.

For all experiments the initial neighborhood range λ_0 was chosen as $n/2$, n being the number of neurons. The neighborhood range $\lambda(t)$ at epoch t was decreased according to $\lambda(t) = \lambda_0 \cdot (0.01/\lambda_0)^{t/t_{\max}}$ (cf. [11]), $t_{\max}$ being the number of epochs. If not indicated otherwise, the number of epochs was 100.

Control experiment

At first, the experiment from [6] for Euclidean data was repeated for RNG as control. Data were sampled from the distribution $(v_1, \ldots, v_d, \prod_{j=1}^{d} \sin(\pi \cdot v_j))$ for $d \in \{1, 2, 3\}$ and uniform $v_i \in [0, 1]$. The number of stimuli was chosen as 2500 for $d = 1$, 5000 for $d = 2$, and 10000 for $d = 3$. We trained RNG for control values $m \in [-1.5, 3.5]$ and step size 0.25. A NG network with 50 neurons has been used. The reported results have been averaged over 20 runs.

The information theoretic quality of the map can be judged by computing the map entropy as reported in Fig. 1. The entropy should be maximum for optimum information transfer, i.e. for $m = 2$ ($d = 1$), $m = 1$ ($d = 2$), and $m = 2/3$ ($d = 3$). As indicated by the arrows, the experimental optima of the curves are closely situated to the expected theoretical values.

Protein Dataset

The evolutionary distance of 226 globin proteins is determined by alignment. These samples originate from different protein families: hemoglobin-α, hemoglobin-β, myoglobin, etc. Here, we distinguish five classes as proposed in [4]: HA (31.86%), HB (31.86%), MY (17.26%), GG/GP (13.27%), and Others 5.75%). Note that the class *Others* combines small classes from the original dataset and represents only a small fraction of the whole dataset.

For the experiment RNG with magnification control parameter $m \in [-1.5, 3.5]$ (step size 0.1) and 50 neurons was trained. The results presented in Fig. 2 are the average over 100 runs. The theoretical optimum $m^* \approx 0.63$ for the Euclidean case as indicated by the arrow in Fig. 2 was derived from the estimated intrinsic dimension $D \approx 3.18$. Note that magnification control by localized learning is

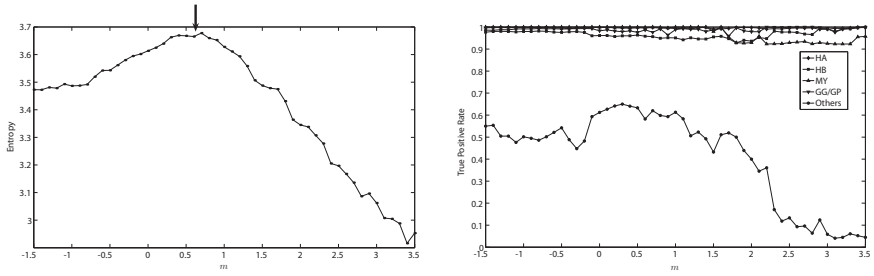

Fig. 2: Protein Dataset – Left: Entropy for different magnification control parameter values m, Right: true positives rates

obviously possible for this non-Euclidean setting. Interestingly, the information theoretic optimum of the curve is closely situated to the Euclidean one.

To demonstrate the magnification effect on exterior (small) classes, the true positives rate for each class is depicted in Fig. 2. Apparently, the classification rate is getting better when focusing on rare events, i.e. for small m.

Chicken Pieces Silhouettes Dataset

The task is to classify 446 silhouettes of chicken pieces into 5 categories (wing, back, drumstick, thigh and back, breast). Data silhouettes are represented as a string of the angles of consecutive tangential line pieces of length 20 and compared using a (rotation invariant) edit distance, where insertions/deletions cost 60, and the angle difference is taken otherwise. We trained a RNG network with magnification control using 50 neurons and control parameter $m \in [-1.5, 3.5]$ (step size 0.1). The average over 100 runs for each different value m was taken.

The arrow in Fig. 2 indicates the theoretical optimum $m^* \approx 0.54$ for the Euclidean case that was derived from the estimated intrinsic dimension $D \approx 3.72$.

Chromosome Images Dataset

The Copenhagen chromosomes database is a benchmark from cytogenetics. featuring 4200 human chromosomes from 22 classes, represented by grey images. These images were transferred to strings of chromosome thickness and compared by alignment. RNG with magnification control has been trained using 80 neurons for control parameter $m \in [-1.5, 3.5]$ (step size 0.25).

The results shown in Fig. 3 present the average over 10 runs for each different value m. The figure shows very smooth control of the map entropy by localized learning. The observed optimum for the considered metric differs from the corresponding optimum $m^* \approx 0.68$ ($D \approx 2.93$) in the Euclidean case.

Cat Cortex Dataset

The Cat Cortex Data Set originates from anatomic studies of cats' brains. A matrix of connection strengths between 65 cortical areas was compiled from literature. For our experiments a preprocessed version of the data set from Haasdonk et al. [4] was used with symmetric matrix which violates the triangle inequality. Note that relational clustering works quite well also in this case of non-metric data. For the experiment, RNG with magnification control has been trained using 12 neurons for control parameter $m \in [-1.5, 3.5]$ (step size 0.1). The results shown in Fig. 3 (right) present the average over 100 runs.

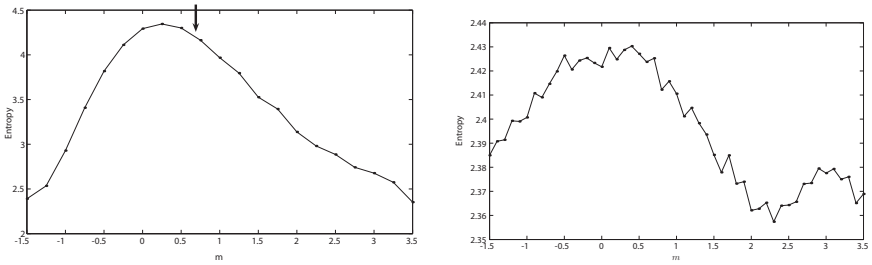

Fig. 3: Chromosome Dataset (left) and CatCortex Dataset (right) – Entropy for different magnification control parameter values m

6 Conclusions

We have extended magnification control by localized learning to Relational Neural Gas, a very powerful extension of NG for general dissimilarity data. The theory transfers directly from standard NG if an Euclidean embedding of data exists; for the general setting of symmetric and nonsingular Δ, convergence can be guaranteed and the dual cost functionis optimized. Experiments demonstrated a very robust and smooth behavior that can be beneficial in practical applications. Thus, magnification control is possible also in the non-Euclidean case using localized learning, although the exact location of the information-theoretic optimum is not known. Depending on the considered metric or matrix Δ, the location of the optimum can change compared to the Euclidean setting as it was shown in the context of Concave/Convex Learning for NG [13].

References

[1] M. Cottrell, B. Hammer, A. Hasenfuss, and T. Villmann (2006), Batch and median neural gas, *Neural Networks*, 19:762-771.

[2] J. Geffroy, Sur l'estimation d'une densité dans un espace métrique, *C.v.Acad.Sci.Paris*, Sér.A 278; 1449-1452, 1974.

[3] P. Grassberger and I. Procaccia, Measuring the strangeness of strange attractors. *Physica D*, **9:**189-208, 1983.

[4] B. Haasdonk and C. Bahlmann, Learning with distance substitution kernels, in *Pattern Recognition - Proc. of the 26th DAGM Symposium*, 2004.

[5] B. Hammer and A. Hasenfuss, Relational Neural Gas. In J. Hertzberg et al., editors, *KI 2007*, pages 190-204.

[6] B. Hammer, A. Hasenfuss, and T. Villmann, Magnification control for batch neural gas, *Neurocomputing*, 70:1225-1234, 2007.

[7] B. Hammer, B. J. Jain, Neural methods for non-standard data, *ESANN'2004*, p.281-292.

[8] A. Hasenfuss and B. Hammer, Relational Topographic Maps. In *IDA'2007*, p.93-105.

[9] A. Jain and E. Merény, Forbidden Magnification? I, *ESANN'2004*.

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

[11] T. Martinetz, S. Berkovich, and K. Schulten, 'Neural gas' network for vector quantization and its application to time series prediction. *IEEE TNN*, 4(4):558–569, 1993.

[12] M. Neuhaus and H. Bunke, Edit distance based kernel functions for structural pattern classification. *Pattern Recognition* 39(10):1852-1863, 2006.

[13] T. Villmann and J. C. Claussen Magnification control in self-organizing maps and neural gas, *Neural Computation*, 18(2):446-469, 2006.

[14] T. Villmann and A. Heinze, Application of magnification control for the neural gas network in a sensorimotor architecture for robot navigation. *SOAVE'2000*, 125–134, 2000.

[15] P. Zador, Asymptotic quantization error of continuous signals and the quantization dimension. *IEEE Transactions on Information Theory*, 28:149–159, 1982.