ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Supporting GNG-based Clustering with Local Input Space Histograms

Jochen Kerdels and Gabriele Peters

University of Hagen - Chair of Human-Computer Interaction
Universitätsstrasse 1, 58097 Hagen - Germany

Abstract. This paper presents an extension to the *growing neural gas* (GNG) algorithm that allows to capture local characteristics of the input space. Using these characteristics clustering schemes based on the GNG network can be improved by discarding uncertain edges of the network and identifying edges that span discontinuous regions of input space. We applied the described approach to different two-dimensional data sets found in the literature and obtained comparable results.

1 Introduction

In recent years the ability to store and process vast amounts of data has increased considerably. One approach to discover structure in such data is the use of methods that employ forms of *unsupervised competitive learning*. The *growing neural gas* (GNG) proposed by Fritzke [1] is such a method. It belongs to the class of *topology representing* networks [2]. In contrast to other methods of unsupervised competitive learning, e.g., the self-organizing map (SOM) [3], the GNG does not use a fixed network topology. Instead, the GNG uses a data-driven growth process to approximate the topology of the input space in form of an induced Delaunay triangulation. However, the individual GNG units can only represent local regions of the input space as convex polyhedrons. Hence, complex structures of the input space, e.g., non-convex clusters, can only be approximated piecewise by a larger number of units.

This piecewise representation of input space structures makes it often difficult to recover the relationship between the network units and the corresponding structures in input space. For instance, determining which set of units in the GNG network corresponds to a single, coherent cluster in the input space can be particularly difficult when several other clusters are very close. To support the reconstruction of such relationships, this paper proposes an extension to the original GNG algorithm by adding a local descriptor to each edge of the GNG network that characterizes the input space structure spanned by the particular edge. We show that this descriptor can be utilized in combination with common clustering techniques in order to identify sets of GNG network units that correspond well to non-convex, difficult to separate clusters in the input space.

The paper is organized as follows: section 2 provides a brief overview of related work that uses the GNG to identify structures in the input space. In section 3 we introduce our new local descriptor and discuss in section 4 how it can support network-based clustering techniques. We present experimental results in section 5 and end with concluding remarks in section 6.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

2 Related Work

There exist several approaches to recover or improve the relationship between the GNG network and the input space structure. One general idea is the use of existing clustering methods to group the GNG network units. For example, Mitsyn and Ososkov [4] examine hierarchical clustering of GNG units using *single linkage* and *Ward's method* as linkage criteria. Although their clustering results look promising, their approach has the drawback that they have to manually identify the "right" level in the cluster hierarchy to obtain an adequate clustering of the input space.

Other approaches use the fact that the GNG has a tendency to generate isolated sub-networks if the corresponding structures in the input space are sufficiently distant from one another, e.g., Canales and Chacón [5] perform a post pruning step after the GNG has approximated the input space to remove units that lie in sparse regions in order to increase the chance of obtaining a number of isolated sub-networks that correspond well to dense clusters in the input space. Ocsa et al. [6] incorporate a similar idea directly into the growing process of the GNG to the effect that sparse regions of the input space are virtually ignored from the outset. A slightly different approach is made by Doherty et al. [7]. They use an unmodified GNG with a relatively large number of units to reduce the necessary distance between clusters to form isolated sub-networks. During the growth process they track the formation of isolated sub-networks within a tree structure to preserve their neighborhood relations. However, the structure of the resulting tree depends heavily on the (manually choosen) time interval in which the tree is updated.

These and similar clustering approaches can be aided by the GNG extension that is described in the following sections. Even in combination with a straightforward clustering scheme based on breadth-first search (section 4) the obtained results are comparable to those of more sophisticated methods (section 5).

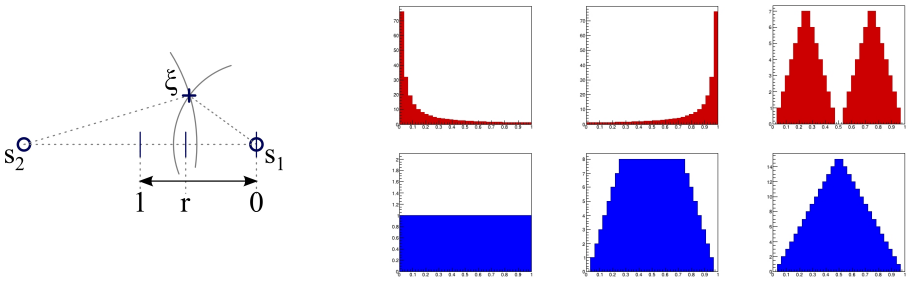
3 Local Input Space Histograms

The growing neural gas uses two mechanisms, edges and accumulated errors, to track how well the network approximates a particular local region of the input space. An edge between two units indicates that the input space between the two units is not empty, and the accumulated error of a unit provides an indication of the relative density of network units with respect to the density of the underlying input space.

We propose to add a small histogram $H = \{h_1, \ldots, h_k\}$, e.g., with $k = 32$ bins, to each edge of the GNG network in order to capture more information about the structure of the input space. Every time two nodes connected by an edge are selected as the two units s_1 and s_2 who are closest to an input ξ, the corresponding histogram is updated with an input distance ratio r defined as:

$$r := \frac{\|s_1 - \xi\| - \|s_2 - \xi\|}{\|s_1 - s_2\|} + 1.$$

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.



(a) (b)

Fig. 1: **(a)** The distance ratio r ranges from 0 to 1 and describes how close the best matching unit s_1 is to the input ξ in relation to the second best matching unit s_2. **(b)** Six of the eleven templates used to identify histograms that represent forms of discontinuous (upper row) and continuous (lower row) input space.

The ratio r is illustrated in figure 1a. It ranges from 0 to 1 and describes how close unit s_1 is to the input ξ relative to unit s_2. Since the histogram H is bound to an edge, it is "used" by two units a and b. In case unit a is the best matching unit the ratio r is inserted in the lower half of the histogram, i.e., in bins h_1 to $h_{k/2}$. Otherwise, the ratio r is inserted in the reversed upper half of the histogram ranging from bin h_k down to bin $h_{k/2}$. The resulting histogram captures the local characteristic of the input space that is spanned by the corresponding edge. Figure 2 shows the local histograms of two edges. One of the edges covers a gap in the input space. The other edge spans a region were the inputs are more equally distributed.

4 Clustering

The additional information that is gathered by the local histograms can be used to support the clustering of GNG network units. The basic algorithm uses a plain breadth-first search to identify network units that are connected by at least one path. Without any extensions, this breadth-first search identifies isolated sub-networks as individual clusters. This straightforward clustering technique can be extended by deciding for every edge, if the edge should actually be traversed.

We propose to use two measures to support this decision: Edges whose information about the input space is uncertain as well as edges that characterize some form of "gap" in the input space should not be traversed. We define the uncertainty u_c of an edge c as the average bin error of its histogram H:

$$u_c := \frac{1}{k'} \sum_{h_i \in H'} \frac{\sqrt{h_i}}{h_i}, \quad k' := |H'|, \; H' := \{h | h \in H \wedge h > 0\}$$

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

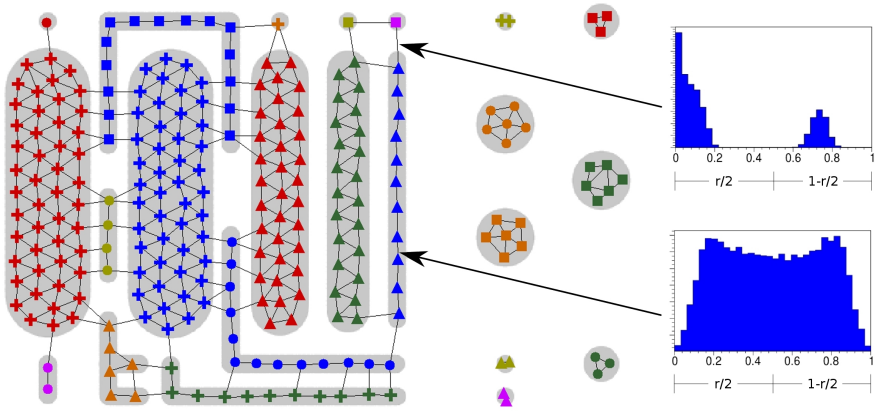


Fig. 2: Clustering of an inhomogeneous input space (light gray) with a GNG network of 250 units resulting in 23 clusters (marked with different shapes and colors at the units' positions). The histograms of two edges (marked by the arrows) are shown on the right.

If the uncertainty u_c of an edge is above a certain threshold, e.g., 10%, the edge will not be traversed.

For the second measure we compare the shape of the edge's histogram with a set of 11 template histograms that represent forms of discontinuous (5) as well as continuous (6) input space (see figure 1b). The templates were determined empirically. If a histogram matches to one of the templates representing discontinuous input space the corresponding edge will not be traversed. In order to calculate the similarity between an edge's histogram and a template we employ the *earth mover's distance* (EMD) [8]. Figuratively speaking, the EMD calculates the minimum amount of "work" needed to transform one histogram into another by moving the bin contents of the first histogram to match the bin contents of the second one. Unfortunately the computation of the EMD has a complexity of $O\left(n^3 \log n\right)$. To avoid this unpleasant property of the EMD we use an approximation of the EMD proposed by Shirdhonkar and Jacobs [9] that is based on the wavelet transform and can be calculated in $O\left(n\right)$.

5 Results

We implemented a GNG extended with the described local input space histograms and applied it to different two-dimensional input spaces. Figure 2 shows the clustering result of an input space with dense regions (light gray) that are partially in close proximity and have non-convex shapes in some cases. In addition, the figure depicts the input space histograms of two edges (black arrows). Although most regions of the input space are connected by edges in their corre-

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

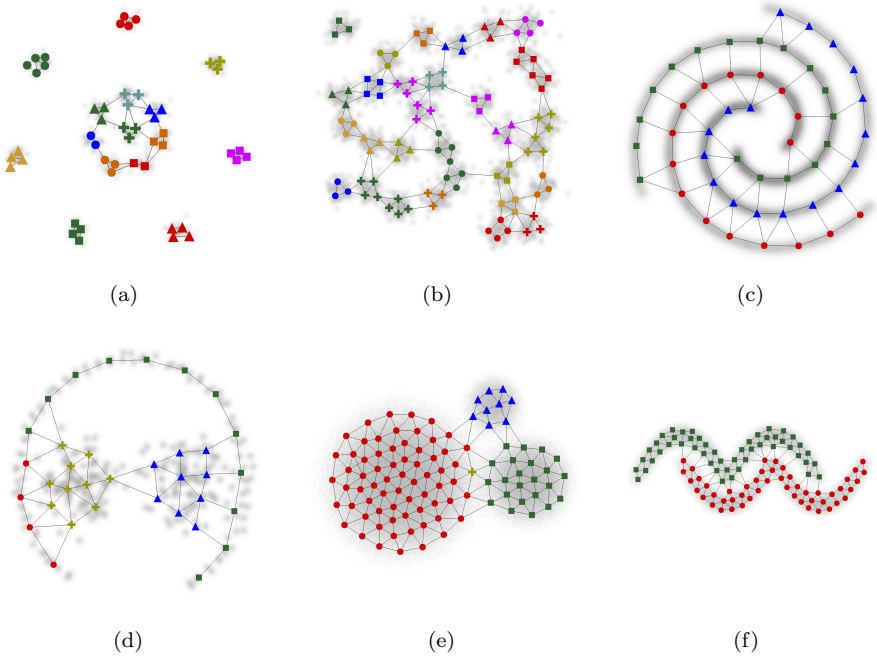


Fig. 3: Application of the described clustering approach to six datasets from the literature. **(a,b)** are based on the R15 and D31 datasets in [10]. **(c,d)** are based on the datasets presented in [11]. **(e,f)** are based on the datasets in [4]. Clusters are marked with different shapes and colors.

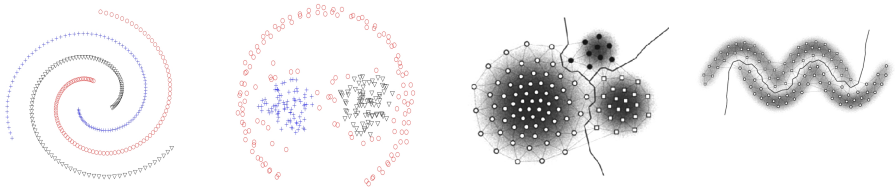
sponding network representation, the two measures derived from the local input space histograms prevent a merging of the particular clusters.

In figure 3 we applied our clustering approach to datasets found in the literature and achieved comparable results (see figure 4). The input spaces in figure 3a and 3b are based on the R15 and D31 datasets in [10], the input spaces in figure 3c and 3d are based on datasets presented in [11] (figure 4a and 4b), and the input spaces in figure 3e and 3f are based on the datasets shown in [4] (figure 4c and 4d). As the datasets of [10] and [11] were given as a set of discrete points, the input spaces were derived by plotting the points as small circles and applying a gaussian filter. In all cases the threshold for the allowed uncertainty of edges was set to 10% and the same set of 11 templates was used.

6 Conclusions

We presented an extension to the GNG algorithm that allows to capture the local characteristics of the input space. We showed how this information can be used to support clustering schemes based on the GNG network by excluding uncertain edges and utilizing a template-based approach to identify discontinuous

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.



(a) (b) (c) (d)

Fig. 4: Examples of clustering results presented in [11] **(a,b)** and [4] **(c,d)**.

regions in input space. The described clustering scheme was applied to different two-dimensional input spaces found in the literature and it yielded comparable results. The current approach uses a small set of hand-crafted templates to identify characteristic regions of the input space. It appears promising to determine if these templates can also be learned in an unsupervised way. Furthermore, the presented idea is, in general, applicable to other prototype based clustering methods, e.g., the standard, non-growing neural gas.

References

[1] Bernd Fritzke. A growing neural gas network learns topologies. In *Advances in Neural Information Processing Systems 7*, pages 625–632. MIT Press, 1995.

[2] Thomas M. Martinetz and Klaus Schulten. Topology representing networks. *Neural Networks*, 7:507–522, 1994.

[3] Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69, 1982.

[4] S.V. Mitsyn and G.A. Ososkov. The growing neural gas and clustering of large amounts of data. *Optical Memory and Neural Networks*, 20(4):260–270, 2011.

[5] Fernando Canales and Max Chacón. Modification of the growing neural gas algorithm for cluster analysis. In Luis Rueda, Domingo Mery, and Josef Kittler, editors, *Progress in Pattern Recognition, Image Analysis and Applications*, volume 4756 of *Lecture Notes in Computer Science*, pages 684–693. Springer Berlin Heidelberg, 2007.

[6] Alexander Ocsa, Carlos Bedregal, and Ernesto Cuadros-Vargas. Db-gng: A constructive self-organizing map based on densilty. In *IJCNN*, pages 1953–1958, 2007.

[7] K.A.J. Doherty, R.G. Adams, and N. Davey. Hierarchical growing neural gas. In Bernardete Ribeiro, RudolfF. Albrecht, Andrej Dobnikar, DavidW. Pearson, and NigelC. Steele, editors, *Adaptive and Natural Computing Algorithms*, pages 140–143. Springer Vienna, 2005.

[8] Y. Rubner, C. Tomasi, and L.J. Guibas. A metric for distributions with applications to image databases. In *Computer Vision, 1998. Sixth International Conference on*, pages 59–66, 1998.

[9] S. Shirdhonkar and D.W. Jacobs. Approximate earth mover's distance in linear time. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, 2008.

[10] Cor J. Veenman, Marcel J. T. Reinders, and Eric Backer. A maximum variance cluster algorithm. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(9):1273–1280, sep 2002.

[11] Hong Chang and Dit-Yan Yeung. Robust path-based spectral clustering. *Pattern Recogn.*, 41(1):191–203, jan 2008.