ESANN'2006 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2006, d-side publi., ISBN 2-930307-06-4.

Data topology visualization for the Self-Organizing Map

Kadim Taşdemir and Erzsébet Merényi *

Rice University - Electrical & Computer Engineering
6100 Main Street, Houston, TX, 77005 - USA

Abstract. The Self-Organizing map (SOM), a powerful method for data mining and cluster extraction, is very useful for processing data of high dimensionality and complexity. Visualization methods present different aspects of the information learned by the SOM to gain insight and guide segmentation of the data. In this work, we propose a new visualization scheme that represents data topology superimposed on the SOM grid, and we show how it helps in the discovery of data structure.

1 Visualization of SOM knowledge

The Self-Organizing Map (SOM) [1] is a widely and successfully used neural paradigm for clustering and data mining. Informative representation of the learned SOM's knowledge greatly aids precise capture of the cluster boundaries. This is especially important for high-dimensional and large data sets with many meaningful clusters such as in remote sensing or medical imagery, which often also have interesting rare clusters to be discovered.

An impressive suite of previous works include the U-matrix [2] and its variants, which are useful when relatively large SOM grid accomodates small data sets with a low number of clusters (e.g., [3], [4], [5]) but, because of averaging of prototype distances over neighbours or thresholding, they tend to miss finer structure in complicated data [6]. Unique approaches such as [7] and gravitational methods (e.g., Adaptive Coordinates [4]) visualize distances between receptive field centres in innovative ways that greatly help manual cluster extraction. Experiments with automated colour assignments aim at qualitative exploration of the approximate cluster structure [8], [9], [10]. We point the reader to [4], [11] for more review. Some earlier methods use the size of the receptive fields of the prototypes for visualization (e.g., [5], [9]), but none exploit the data topology. Visualization of the mapping of samples, adjacent in data space, to different SOM prototypes is useful when prototypes outnumber data samples [12]. When data samples are plenty, adjacent samples mapped to different prototypes are only the ones at the boundaries of the Voronoi polyhedra, causing the visualization to ignore a lot of helpful mapping information. We visualize the data topology on the SOM grid, showing topology violations and effectively aiding in detailed cluster capture including fine structure in large real data with many clusters of widely varying statistics.

*This work was partially supported by grant NNG05GA94G from the Applied Information Systems Research Program, NASA, Science Mission Directorate. Figures are in colour, request colour copy by email: tasdemir@rice.edu, erzsebet@rice.edu

2 Topology representation through "connectivity matrix"

In a trained SOM, the weights of the processing elements (PEs) become prototype vectors, w_i, of the data samples. Here, i denotes the location of prototype w_i in the SOM grid. Each w_i has a receptive field, RF_i, for which it is the best matching unit (BMU). The second BMU is also important since the prototypes are adapted by cooperation of the winner and its neighbours. Their relations form a data distribution within the receptive field, which shows the similarity of w_i to the w_j's that are adjacent to w_i. To represent the data distribution within the receptive fields, we define a cumulative adjacency matrix, $CADJ$.

Definition 1: *Let CADJ be an $M \times M$ matrix where M is the number of all SOM prototypes. The cumulative adjacency, $CADJ(i,j)$, of two prototypes w_i and w_j, is the number of data samples for which w_i is the BMU and w_j is the second BMU.* By this definition, $|RF_i| = \sum_{k=1}^{M} CADJ(i,k)$.

Definition 2: *The level of connectedness of two prototypes w_i and w_j is*

$$CONN(i,j) = CONN(j,i) = CADJ(i,j) + CADJ(j,i). \quad (1)$$

where CONN called connectivity strength matrix and $CONN(i,j)$ is an element of it. By definition, CONN is symmetric and it is the weighted analogue of the adjacency matrix obtained by induced Delaunay triangulation defined in [13]. It shows how strongly two prototypes are connected in data space. CONN is a sparse matrix because of the forced 2-d grid placement of the prototypes on the SOM. Ideally, the SOM is a topology preserving mapping, thus only the immediate neighbours in the SOM should be connected in data space. However, because of noise, outliers, data complexity or badly formed SOM, connections may exist between prototypes which are not immediate neighbours in the SOM.

3 Similarity visualization by connectivity matrix

We visualize CONN by connecting pairs of prototypes w_i and w_j in the SOM grid with lines of various widths and colours for $CONN(i,j) > 0$ (Fig. 1). The line width signifies the connectivity strength to indicate global importance of the connection while the line colour shows the ranking of connectivity strengths among all connections of w_i to represent local importance. As thicker lines block visualization of other connections, the representation of different connectivity strengths with different line widths is limited and should be governed by the specific data and application. Here, we use a four level binning based on the mean (μ) and variance (σ) of the strengths of all connections between prototypes:

$$line\ width(i,j) = \begin{cases} 1 & \mu - \sigma > CONN(i,j) > 0 \\ 2 & \mu > CONN(i,j) > \mu - \sigma \\ 3 & \mu + \sigma > CONN(i,j) > \mu \\ 4 & CONN(i,j) > \mu + \sigma \end{cases} \quad (2)$$

We arbitrarily call a connection with *line width* $= 1$ "weak", and a connection with *line width* > 1 "strong", which provides sufficient resolution for cluster capture for the cases we present here.

Visualization of CONN shows the relations of the prototypes in the data space superimposed on the SOM grid, highlighting topology violations. Fig. 1 is

an example for a 20 x 20 SOM of a 6-band image with 11 classes (the data is described in the next section). The length of violating connections, i.e., the distance between the locations of prototypes in the SOM, with *line length* > 1, shows whether topology violations are local or global. Low strength (*line width* $= 1$) usually indicates outliers or noise while greater strengths are due to data complexity or badly formed SOM. To mitigate the visually obscuring effect of topology violating connections, we can eliminate low strength global violations. Strong violations (thick lines) are formed because of the data complexity or badly formed SOM and they should be evaluated after elimination.

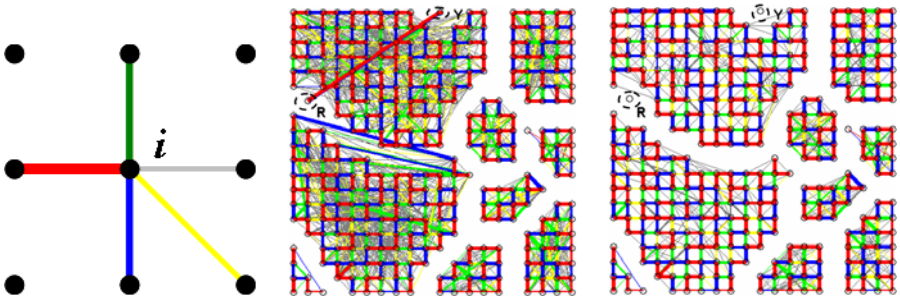

Fig. 1: Visualization of connectivity matrix CONN **Left:** An example of connections for a prototype (center node i). A line is drawn between two nodes if they are adjacent in the data space according to Delaunay triangulation. The line width shows the connectivity strength. The strongest-to-weakest connections of prototype w_i are coded as red, blue, green, yellow and grey scales, respectively. **Middle:** CONN visualization for a 20 x 20 SOM of 6-d, 11-class data set (see Fig. 2 for data). All connections ($CONN(i,j) > 0$) are drawn. The white gaps correspond to dead prototypes (see Fig. 2, left, for ground truth). The connections between the prototypes that are not neighbours in the SOM lattice show topology violations. Most violations are local or within clusters. A strong global violation (thick red line) between two prototypes exists which should be evaluated. This strong global violation connects two small clusters, Y and R. Each is represented by one prototype, and separated strongly from other groups of prototypes. This and the low hit count of the respective single prototypes suggest rare clusters. Since they have very distinct signatures (Fig. 2), we evaluate them as two different clusters in spite of the strong connecting line, and conclude that they are misordered in this SOM. The remaining signature groups are clearly separated because topology violations remain within the respective clusters. **Right:** Connections with *length* > 3 are eliminated.

4 Applications and discussion

We show the applicability of our CONN visualization on 3 data sets. The first is a 6-band 128 x 128 pixel image with 11 classes 3 of which, R, T and Y, are rare and significantly different from the rest. Fig. 2 shows the image and the mean signatures of the classes. The CONN visualization, in Fig. 1, displays the local and global topology violations. All violations except for the rare classes R and Y (connected by a thick red line in Fig. 1, Middle) are within the respective clusters. After careful assessment of the characteristics of these small clusters (as explained Fig. 1), we eliminate the connection to get two rare clusters R and Y. All other clusters are well separated by our visualization.

The second data set is a 6-band 128 x 128 pixel image containing 20 classes 4 of which are rare. 8 classes are in common with the previous set. In this data set,

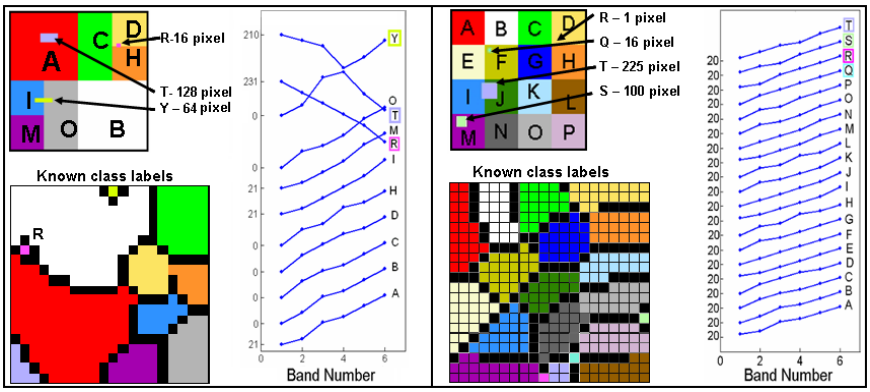

Fig. 2: Two 6-band image data sets: spatial distributions of classes in the images; their known class labels overlain on the SOMs; and mean signatures of classes, offset for clarity. Individual data samples were generated by adding 10% Gaussian noise to the mean signatures. **Left:** The 11-class data. Note that the 16-pixel class R has a hot pink (not red) color. **Right:** The 20-class data set. The rare classes R and T are different from the classes with same labels in the 11-class data set while the other 8 are the same.

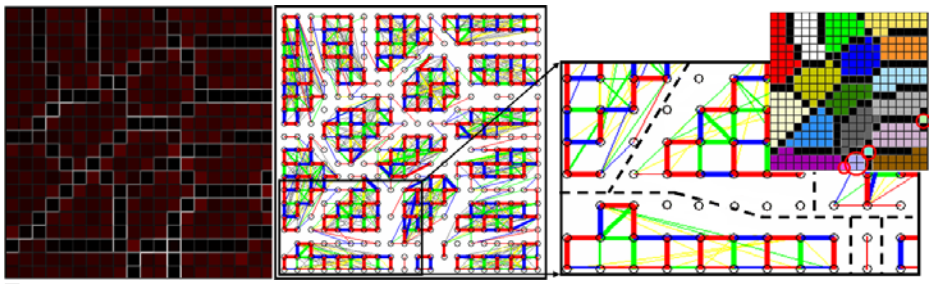

Fig. 3: **Left:** Modified U-matrix of the 20 x 20 SOM for the 20-class data set. Hit counts are indicated by the intensity of the red colour of the cells. White (high) fences indicate strong dissimilarity of neighboring prototypes. **Middle:** CONN visualization of the same SOM. In the U-matrix, prototypes are represented by the centres of the grid cells while in the CONN the prototypes are at the junctions of the connections. Here, thin connections mean weak similarity. Topology violations are mostly within classes. Clusters are separated by weak connections. **Right:** Semi-manual clustering based on CONN visualization. As an example, we show the extraction of the purple (M), light yellow (E) and medium blue (I) clusters. We separate them by cutting the weak connections between the cluster boundaries. To see which connections are eliminated, we leave the nodes that are disconnected from clusters. All classes are correctly identified this way, including the rare ones (in solid color circles).

the rare classes have mean signatures similar to the rest. Fig. 2 shows the layout and the mean signatures of the classes. A comparison of the U-matrix and CONN representations of the SOM is given in Fig. 3. We identify the clusters from the CONN visualization by evaluating the width and colour of connections, and the number of neighbours connected to the PEs across cluster boundaries. First the weakest line is eliminated. From links with the same width, the lowest ranking lines are cut (see Fig. 3). This procedure correctly extracted all classes including the rare ones. For these simple data, extraction of cluster boundaries based either on U-matrix or CONN works well. However, for complicated data the CONN visualization gives more assistance since it represents the data topology

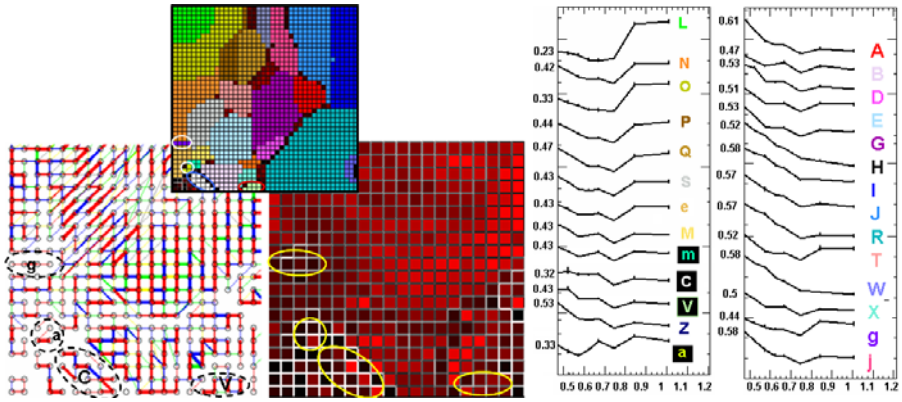

Fig. 4: **Left:** Comparison of CONN visualization (left) and U-matrix (right) for the 40x40 SOM of Ocean City. The lower left quadrants (20x20) are shown for both cases. In the CONN, prototypes are at the junctions of the connections while in the U-matrix, prototypes are represented by the centres of the grid cells. For CONN visualization, the weak global violations were excluded. In the middle inset, labels of the 27 clusters extracted from the CONN visualization are shown. Ovals in the inset and U-matrix, and dashed ovals in the CONN visualization show our capture of the rare clusters (C, V, a and g) which were extracted in previous work [6]. The boundaries of these rare clusters are clearly visible in the SOM by CONN visualization while there are high fences within these clusters in U-matrix representation. **Right:** Mean signatures of selected clusters, offset for clarity. The standard deviations (vertical bars on the signatures) are very small, an indication of the clustering quality.

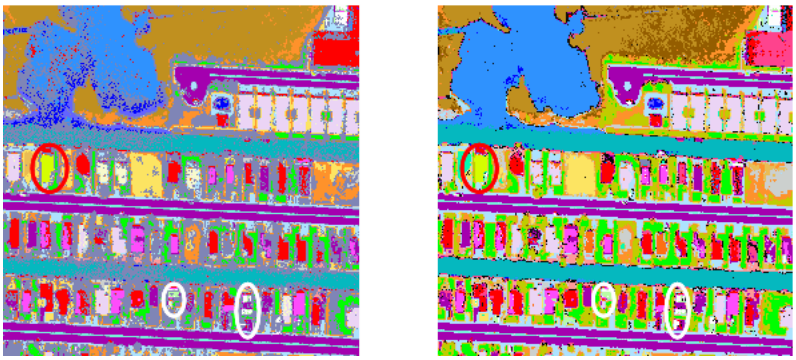

Fig. 5: Comparison of the extracted cluster maps of Ocean City. Only the top right quadrants (256 x 256 pixels) are shown due to space constraints. **Left:** Earlier cluster map extracted by using modified U-matrix (see [6] for details). Red and white ovals show the locations of rare clusters (C, V and a in Fig. 4) **Right:** Clusters extracted from CONN visualization. The agreement between the two cluster map is very good. Here, the majority of the pixels are assigned to clusters, which produces more appearances of some colours such as turquoise and green. We also easily capture the formerly identified rare clusters (shown in the ovals). See Fig. 4 for their labels and locations in the SOM. We thank Dr. Bea Csathó, Ohio State University, Byrd Polar Institute, for the Ocean City data and ground truth.

in relation to the SOM grid, and makes delineating the clusters easier and faster.

The third data set is a real remote sensing spectral image of Ocean City, Maryland, comprising 512 x 512 pixels in 8 spectral bands (each pixel is represented by an 8-d feature vector called spectrum). We use a previous semi-manual

clustering based on modified U-matrix representation (see [6] for details) for performance comparison. Our cluster map extracted from CONN visualization shown in Fig. 4 has a general agreement with the earlier cluster map (Fig. 5). By the clear partition of the CONN visualization, we easily capture the rare clusters (C, V, a and g in Fig. 4 and Fig. 5) that were formerly extracted by [6].

5 Conclusion

Our CONN visualization for trained SOMs is a new, promising method. By representing data topology on the SOM grid, it shows data distribution among the connected prototypes and gives more insight about the similarity of the prototypes than previous representations. In particular, fine structure in complicated data, which often remains obscure in other visualizations, becomes clear. In addition, topology violations of the mapping, and the severity of those violations are indicated superimposed on the SOM. CONN visualization can potentially aid in the automation of clustering and thus be a powerful tool in data mining.

References

[1] T. Kohonen. *Self-Organizing Maps*. Springer-Verlag Berlin Heidelberg, 2nd edition, 1997.

[2] A. Ultsch. Self-organizing neural networks for visualization and classification. In O.Opitz B. Lausen and R. Klar, editors, *Information and Classification-Concepts, Methods and Applications*, pages 307–313. Springer Verlag, Berlin, 1993.

[3] M.A. Kraaijveld, J. Mao, and A.K. Jain. A nonlinear projection method based on Kohonen's topology preserving maps. *IEEE Trans. on Neural Networks*, 6(3):548–559, 1995.

[4] D. Merkl and A. Rauber. Alternative ways for cluster visualization in Self-Organizing Maps. In *Proc. 1st Workshop on Self-Organizing Maps (WSOM97)*, 1997.

[5] A. Ultsch. Maps for the visualization of high-dimensional data spaces. In *Proc. 4th Workshop on Self-Organizing Maps (WSOM03)*, volume 3, pages 225–230, 2003.

[6] E. Merényi and A. Jain. Forbidden magnification? II. In *Proc. 12th European Symposium on Artificial Neural Networks (ESANN), Bruges, Belgium, D-Facto*, pages 57–62, 2004.

[7] M. Cottrell and E. de Bodt. A Kohonen map representation to avoid misleading interpretations. In *Proc. European Symposium on Artificial Neural Networks (ESANN), Bruges, Belgium, D-Facto*, pages 103–110, 1996.

[8] J. Himberg. A SOM based cluster visualization and its application for false colouring. In *Proc. IEEE-INNS-ENNS International Joint Conf. on Neural Networks, Como, Italy*, volume 3, pages 587–592, 2000.

[9] S. Kaski, T. Kohonen, and J. Venna. Tips for SOM processing and colourcoding of maps. In T. Kohonen G. Deboeck, editor, *Visual Explorations in Finance Using Self-Organizing Maps*. London, 1998.

[10] T. Villmann and E. Merényi. Extensions and modifications of the Kohonen SOM and applications in remote sensing image analysis. In L. C. Jain U. Seiffert, editor, *Self-Organizing Maps: Recent Advances and Applications*, pages 121–145. Springer-Verlag, 2001.

[11] J. Vesanto. SOM-based data visualization methods. *Intelligent Data Analysis*, 3(2):111–126, 1999.

[12] G. Polzlbauer, A. Rauber, and M. Dittenbach. Advanced visualization techniques for self-organizing maps with graph-based methods. In *Proc. Intl. Symp. on Neural Networks (ISSN05)*, pages 75–80, 2005.

[13] T. Martinetz and K. Schulten. Topology representing networks. *Neural Networks*, 7(3):507–522, 1993.