Topological Correlation

K.A.J. Doherty, R.G. Adams and and N. Davey

University of Hertfordshire, Department of Computer Science
College Lane, Hatfield, Hertfordshire, UK

Abstract. Quantifying the success of the topographic preservation achieved with a neural map is difficult. In this paper we present *Topological Correlation*, T_c, a method that assesses the degree of topographic preservation achieved based on the linear correlation between the topological distances in the neural map, and the topological distances in the induced Delaunay triangulation of the network nodes. In contrast to previous indices, T_c has been explicitly devised to assess the topographic preservation within neural maps composed of many sub-graph structures. The T_c index is bounded, and unequivocally identifies a perfect mapping, but more importantly, it provides the ability to quantitatively compare less than successful mappings. The T_c index also provides an indication of the maximum number of nodes to use within the neural map.

1 Introduction

Topographic clustering algorithms are grouped under the general label of *neural maps*, and all use graph structures to build a representation of the input space: a well-known example is the SOFM [1]. We are interested in neural maps that designate clusters in the data with discrete sub-graph structure in the neural map, such as the models generated by the Growing Neural Gas (GNG) [2] and Growing Cell Structures (GCS) [3] algorithms. A review of the literature showed that none of the current cluster validity indices were suitable for determining the success of the clustering produced with these algorithms, and this motivated the work presented in this paper.

2 The Measurement of Topographic Preservation

The literature is rich in definitions of topographic preservation measures, e.g., see [4, 5, 6], and many others. The measures proposed in the literature use various combinations of *metric*, *rank* and *topological* measures of similarity. These measures were shaped by the differing interpretations that researchers apply to defining the topography of a neural map, and these approaches contain interesting ideas. However, most of these measures focus on the problem of determining the most appropriate dimensionality of a regular lattice of nodes, and whilst the use of such a lattice of nodes is popular, there are topographic mapping techniques that are not restricted to either a prespecified or fixed dimensionality. Moreover, the topographic preservation metrics assume that the neural map is a single graph, that has no sub-graph structure, and none explicitly specify how to measure distances in the neural map between the disconnected sub-graphs.

Some of these measures only give an indication of topographic preservation errors within immediate neighbours, and take no account of larger topographic preservation errors which may limit their usefulness in identifying gross violations in topographic preservation. The measure of topographic preservation we present in the next section, successfully addresses these problems.

3 Topological Correlation

We now introduce our measure of topographic preservation, the *Topological Correlation* index, T_c. The concept of topological neighbourhood (i.e., adjacency) is central, in our opinion, to what constitutes a natural cluster. The measurement of distance by considering topological relationships between those Voronoi polyhedra that contain data points (the masked Voronoi polyhedra [7]), rather than the full Voronoi polyhedra, has an intuitive appeal, as the neighbourhood relationships between network nodes are derived through the data distribution.

The T_c index provides an quantitative method for the evaluation of the success of a topographic mapping. It achieves this by calculating the linear correlation between two distances. The first distance d_V is the topological distance (i.e., path length) in the induced Delaunay triangulation [7] of the positional vectors in the input space. The second distance d_G is the topological distance in the the network graph. Hence, T_c is measuring the correlation between two measures of neighbourhood adjacency. The T_c index is given by:

$$T_c = \frac{\sum_{i=2}^{n}\sum_{j=1}^{i-1}(d_{G(ij)} - \overline{d}_G)(d_{V(ij)} - \overline{d}_V)}{\sqrt{\left(\sum_{i=2}^{n}\sum_{j=1}^{i-1}(d_{G(ij)} - \overline{d}_G)^2\right)\left(\sum_{i=2}^{n}\sum_{j=1}^{i-1}(d_{V(ij)} - \overline{d}_V)^2\right)}} \tag{1}$$

where $\overline{d}_G$ and $\overline{d}_V$ are the mean of the entries in the lower half of the d_G and d_V distance matrices, given by $\overline{d}_G = \frac{\sum_{i=2}^{n}\sum_{j=1}^{i-1} d_{G(ij)}}{n(n-1)/2}$ and $\overline{d}_V = \frac{\sum_{i=2}^{n}\sum_{j=1}^{i-1} d_{V(ij)}}{n(n-1)/2}$ respectively. Furthermore, $d_{G(ij)}$ and $d_{V(ij)}$ are the minimum path lengths between two graph nodes i and j, in the network graph and the ideal induced Delaunay triangulation of the network nodes. The use of minimum path length as the measure of topographic similarity allows the index to indicate minor deviations in topological preservation. By using zero for either or both d_G and d_V where no path exists, it provides the ability to highlight regions of the graph where paths exist between sub-graph structures where they should not, and vice-versa. If there is no path between i and j, then $d_{(ij)}$ is zero, and thus $d_{(ij)}$ is a *pseudometric* as it fails to satisfy the *identity of indiscernables* axiom of metricity. The T_c index is bounded in the range $T_c \in [-1, 1]$, and is interpreted as other correlation coefficients, viz, $T_c = 1$ is indicative of a perfect positive linear corre-

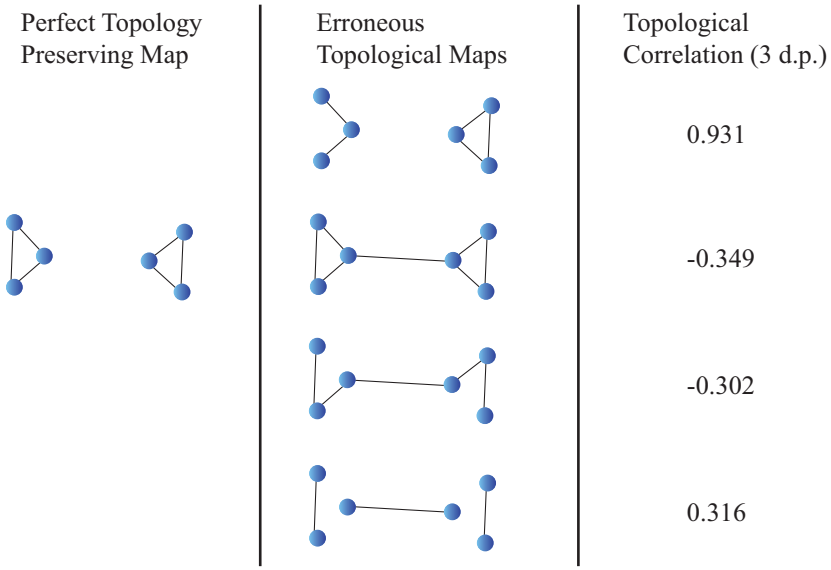

Fig. 1: Examples of the measurement of topological correlation, T_c. The graph in the left column is an example perfect topological preserving map against which some erroneous topological maps (center column) are measured. The T_c between the perfectly topological preserving map and the erroneous topological maps is shown in the right column.

lation, $T_c = -1$ is indicative of a perfect negative linear correlation, and $T_c = 0$ indicates that no linear correlation exists.

A simple example of the application of the T_c index is shown in Fig. 1. It is clear from this example that minor topological preservation errors such that the correct large scale sub-graph structures are identified, but which may still contain inappropriate edge structure (e.g., the upper graph in the center column of Fig. 1) are indicated with a small deviation from a perfect correlation. But large scale errors, such that the correct sub-graph structure is lost (e.g., the remaining graphs in the center column of Fig. 1), produces much a larger deviation from a perfect correlation. Used in isolation, the T_c index does not provide a measure of clustering quality. What it *does* provide is the ability to quantify the suitability of a network graph structure in relation to the topologically ideal graph for a given set of data. When combined with a measure of the *quality* of the spread of the network nodes, we suggest that the quality of a clustering scheme can be evaluated. This combination of T_c and quality of the network node spread *could* be combined in some ad-hoc fashion to quantify the success of a topographic mapping, but we take the view (as do [8]) that an investigator can draw their own conclusions from the two separate results.

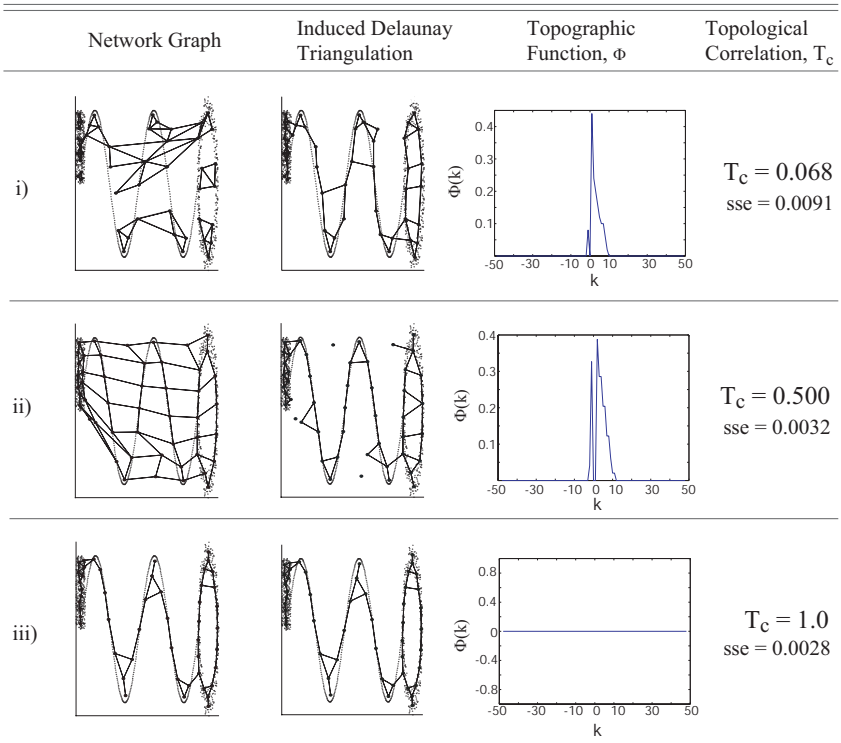

Fig. 2: A comparison of the Topographic Function, Φ, and the Topological Correlation, T_c, on a complex data-manifold. In i), the results of a GCS network can be seen. The network graph contains 6 sub-graphs, whereas the induced Delaunay triangulation indicates that a continuous graph is appropriate for these positional vectors. The Topographic Function indicates errors by taking a non-zero value in its plot. For this network graph it shows a large number of topographic errors in $\Phi(+k)$, and a smaller number of errors in $\Phi(-k)$. The T_c index of 0.068 reflects the inappropriate structure of the network sub-graphs. In ii), the results of clustering with a 7x7 SOFM can be seen. The induced Delaunay triangulation for these positional vectors suggests that a single graph is appropriate (which, of course, the SOFM naturally fulfills), but the Topographic Function shows errors in both $\Phi(+k)$ and $\Phi(-k)$. The T_c index of 0.500 reflects the large number of topological errors. In iii), the results of clustering with a GNG network with a maximum size of 49 nodes can be seen. The network graph is an exact match for the induced Delaunay triangulation, and this is reflected in both $\Phi(k)$ that shows no error for all values of k, and in T_c which reports a perfect positive linear correlation.

4 Evaluation

This section reports the findings of an evaluation of the performance of the Topological Correlation index against the Topographic Function [6]. The Topographic Function also measures similarity relationships within the data-manifold with a topological neighbourhood based on the induced Delaunay triangulation.

Using the GNG, GCS and SOFM algorithms, we assessed the performance of T_c on a complex data-manifold consisting of 1080 elements arranged across three distinct, but connected regions, embedded in $\Re^2$. The first region is a 2-dimensional rectangular uniform distribution. This leads into a 1-dimensional sine wave that extends for approximately 4π radians. This leads into a elliptical path of data that has been extended into two dimensions with the addition of noise. The results are shown in Fig. 2, and are described in the figure caption. It is interesting to note that sub-figure iii), illustrates the equivalence of the Topographic Function and the Topological Correlation in the case of a perfect topology preserving mapping.

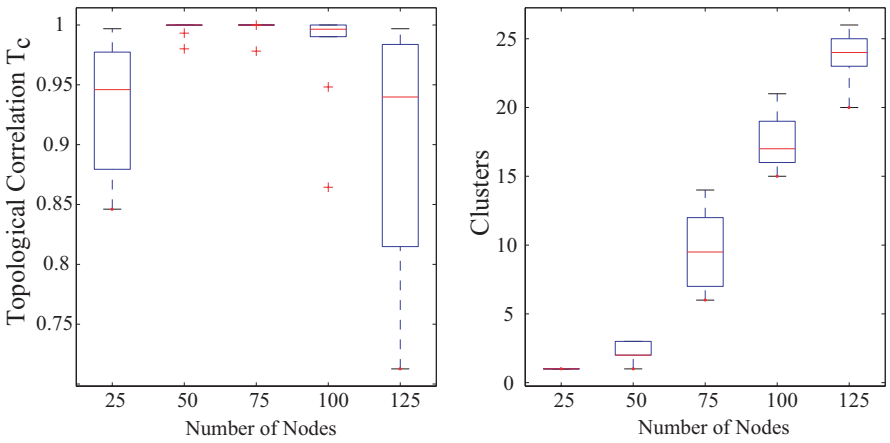

Fig. 3: The T_c index of GNG clustering of a 140 element image data set with varying maximum network sizes. In this case, the GNG networks were unable to satisfactorily distribute the nodes of networks larger than 100 nodes, and we consider that these data form approximately 17 natural clusters.

A very important property of the T_c index is its ability to indicate whether the neural map is composed of too many network nodes. Many neural map algorithms either require the number of nodes to be specified prior to training, or dynamically insert nodes to reduce the estimation criterion. The upper bound for the number of network nodes should be equal to the number of elements in the data set, and for clustering it should (sensibly) be less. As the network size approaches this upper bound, achieving full coincidence of the data and network positional vectors is difficult, and—even with soft competition and other node

distribution techniques—it is very likely that there will be regions of the input data that have too few representative nodes, and there will be regions of the input space that have too many representative nodes. The regions of the input data that have too many representative nodes cause topological errors, that typically manifest as small scale path-length violations. Fig. 3 shows the T_c results for the GNG clustering of a 140–element image data set. This box-and-whisker plot clearly shows that the network size should be limited to less than 100 nodes to produce clusters with a high degree of topological correlation.

We have successfully applied T_c to the task of measuring the quality of clustering a bitmap image data set and a time-series of stock market share closing prices.

5 Conclusions

The Topological Correlation index, T_c, is a new method for determining the degree of topographic preservation in neural maps. The T_c index is bounded, and unequivocally identifies a topology preserving mapping in neural maps that are composed of either a single graph or many sub-graph structures. A very important property of the T_c index, is its ability to indicate the maximum number of network nodes with which a topology preserving neural map can be generated.

References

[1] T. Kohonen. The Self-Organizing Map. *Proceedings of the IEEE*, 78(9):1464–1480, September 1990.

[2] B. Fritzke. A Growing Neural Gas Network Learns Topologies. In G. Tesauro, D.S. Touretzky, and T.K. Leen, editors, *Advances in Neural Processing Systems 7 (NIPS'94)*, pages 625–632, Cambridge, 1995. MIT Press.

[3] B. Fritzke. Growing Cell Structures - a self-organising network for unsupervised and supervised learning. *Neural Networks*, 7(9):1441–1460, 1994.

[4] H.-U. Bauer and K. Pawelzik. Quantifying the Neighbourhood Preservation of Self-Organizing Maps. *IEEE Trans. on Neural Networks*, 3(4), 1992.

[5] J. C. Bezdek and N. R. Pal. An index of topological preservation and its application to self-organizing feature maps. In *Proc. Int. Joint Conf. on Neural Networks*, 1993.

[6] T. Villmann, R. Der, and T. M. Martinetz. Topology Preservation in Self-Organizing Feature Maps: Exact Definition and Measurement. *IEEE Trans. on Neural Networks*, 8(2):256–266, March 1997.

[7] T. M. Martinetz and K. J. Schulten. Topology representing networks. *Neural Networks*, 7(3):507–522, 1994.

[8] H. U. Bauer, M. Herrmann, and T. Villmann. Neural maps and topographic vector quantization. *Neural Networks*, 12:659–676, 1999.