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

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:

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

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-

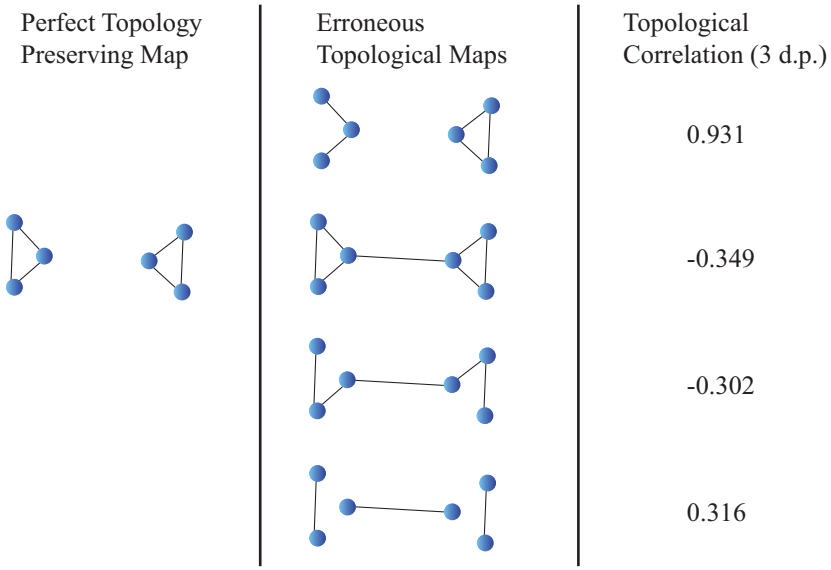

Fig. 1: Examples of the measurement of topological correlation, T_c. The graph in the left column is an example perfect topological preserving map against which some erroneous topological maps (center column) are measured. The T_c between the perfectly topological preserving map and the erroneous topological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