Growing Self-Organizing Maps for Data Analysis

Soledad Delgado
Technical University of Madrid, Spain

Consuelo Gonzalo
Technical University of Madrid, Spain

Estíbaliz Martínez
Technical University of Madrid, Spain

Águeda Arquero
Technical University of Madrid, Spain

INTRODUCTION

Currently, there exist many research areas that produce large multivariable datasets that are difficult to visualize in order to extract useful information. Kohonen self-organizing maps have been used successfully in the visualization and analysis of multidimensional data. In this work, a projection technique that compresses multidimensional datasets into two dimensional space using growing self-organizing maps is described. With this embedding scheme, traditional Kohonen visualization methods have been implemented using growing cell structures networks. New graphical map displays have been compared with Kohonen graphs using two groups of simulated data and one group of real multidimensional data selected from a satellite scene.

BACKGROUND

Data mining first stage usually consist of building simplified global overviews of data sets, generally in graphical form (Tukey, 1977). At present, the huge amount of information and its multidimensional nature complicates the possibility to employ direct graphic representation techniques. Self-Organizing Maps (Kohonen, 1982) fit well in the exploratory data analysis since its principal purpose is the visualization and the analysis of nonlinear relations between multidimensional data (Rossi, 2006). In this sense, a great variety of Kohonen's SOM visualization techniques (Kohonen, 2001) (Ultsch & Siemon, 1990) (Kraaijveld, Mao & Jain, 1995) (Merlk & Rauber, 1997) (Rubio & Giménez 2003) (Vesanto, 1999), and some automatic map analysis (Franzmeier, Witkowski & Rückert 2005) have been proposed.

In Kohonen's SOM the network structure has to be specified in advance and remains static during the training process. The choice of an inappropriate network structure can degrade the performance of the network. Some growing self-organizing maps have been implemented to avoid this disadvantage. In (Fritzke, 1994), Fritzke proposed the Growing Cell Structures (GCS) model, with a fixed dimensionality associated to the output map. In (Fritzke, 1995), the Growing Neural Gas is exposed, a new SOM model that learns topology relations. Even though the GNG networks get best grade of topology preservation than GCS networks, due to the multidimensional nature of the output map it cannot be used to generate graphical map displays in the plane. However, using the GCS model it is possible to create networks with a fixed dimensionality lower or equal than 3 that can be projected in a plane (Fritzke, 1994). GCS model, without removal of cells, has been used to compress biomedical multidimensional data sets to be displayed as two-dimensional colour images (Walker, Cross & Harrison, 1999).

GROWING CELL STRUCTURES VISUALIZATION

This work studies the GCS networks to obtain an embedding method to project the bi-dimensional output

map, with the aim of generating several graphic map displays for the exploratory data analysis during and after the self-organization process.

Growing Cell Structures

The visualization methods presented in this work are based on self-organizing map architecture and learning process of Fritzke's Growing Cell Structures (GCS) network (Fritzke, 1994). GCS network architecture consists of connected units forming k-dimensional hypertetrahedron structures linked between them. The interconnection scheme defines the neighbourhood relationships. During the learning process, new units are added and superfluous ones are removed, but these modifications are performed in such way that the original architecture structure is maintained.

The training algorithm is an iterative process that performs a non-linear projection of the input data over the output map, trying to preserve the topology of the original data distribution. The self-organization process of the GCS networks is similar that in Kohonen's model. For each input signal the best matching unit (*bmu*) is determined, and *bmu* and its direct neighbour's synaptic vectors are modified. In GCS networks each neuron has associated a resource, which can represent the number of input signals received by the neuron, or the summed quantization error caused by the neuron. In every adaptation step the resource of the *bmu* is conveniently modified. A new neuron is inserted between the unit with highest resource, q, and its direct neighbour with the most different reference vector, f, after a fixed number of adaptation steps. The new unit synaptic vector is interpolated from the synaptic vectors of q and f, and the resources values of q and f are redistributed too. In addition, neighbouring connections are modified in order to ensure the output architecture structure. Once all the training vectors have been processed a fixed number of times (epoch), the neurons whose reference vectors fall into regions with a very low probability density are removed. To guarantee the architecture structure some neighbouring connections are modified too. Relative normalized probability density estimation value proposed in (Delgado, 2004) has been used in this work to determine the units to be removed. This value provides better interpretation of some training parameters, improving the removal of cells and the topology preserving of the network. Several separated meshes could appear in the output map when superfluous units are removed.

When the growing self-organization process finishes, the synaptic vectors of the output units along with the neighbouring connections can be used to analyze different input space properties visually.

Network Visualization: Constructing the Topographic Map

The ability to project high-dimensional input data onto a low-dimensional grid is an important property of Kohonen feature maps. By drawing the output map over a plane it will be possible to visualize complex data and discover properties or relations of the input vector space not expected in advance. Output layer of Kohonen feature maps can be printed on a plane easily, painting a rectangular grid, where each cell represents an output neuron and neighbour cells correspond to neighbour output units.

GCS networks have less regular output unit connections than Kohonen ones. When k=2 architecture factor is used, the GCS output layer is organized in groups of interconnected triangles. In spite of bi-dimensional nature of these meshes, it is not obvious how to embed this structure into the plane in order to visualize it. In (Fritzke, 1994), Fritzke proposed a physical model to construct the bi-dimensional embedding during the self-organization process of the GCS network. Each output neuron is modelled by a disc, with diameter d, made of elastic material. Two discs with distance d between centres touch each other, and two discs with distance smaller than d repeal each other. Each neighbourhood connection is modelled as an elastic string. Two discs connected but not touching are pulled each other. Finally, all discs are positively charged and repeal each other. Using this model, the bi-dimensional topographic coordinates of each output neuron can be obtained, and thus, the bi-dimensional output meshes can be printed on a plane.

In order to obtain the output units bi-dimensional coordinates of the topographic map (for k=2), a slightly modified version of this physical model has been used in this contribution. At the beginning of the training process, the initial three output neurons are placed in the plane in a triangle form. Each time a new neuron is inserted, its position in the plane is located exactly halfway of the position of the two neighbouring neurons between which it has been inserted. After this, attraction