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 multi-dimensional 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 multi-dimensional 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 and repulsion forces are calculated for every output neuron and its positions are consequently moved. The attraction force of a unit is calculated as the sum of individual attraction forces that all neighbouring connections exercise over it. Attraction force between two neighbouring neurons *i* and *j*, with p_i and p_j coordinates in the plane, and Euclidean distance *e*, is calculated as $(e-d)/2$ if $e \geq d$, and 0 otherwise. The repelling force of a unit is calculated as the sum of individual repulsion forces that all no-neighbouring output neurons exercise over it. Repelling force between two no-neighbouring neurons *i* and *j* is calculated as $d/5$ if $2d<e\leq 3d$, $d/2$ if $d<e\leq 2d$, d if $0<e\leq d$, and 0 otherwise. There exist three basic differences between the embedding model used in this work and the Fritzke's one. First, repelling force is only calculated with no-neighbouring units. Second, attracting force between two neurons *i* and *j* is multiplied by the distance normalization $((p_j-p_i)/e)$ and by the attraction factor 0.1 (instead of 1). Last, repelling force between two neurons *i* and *j* is multiplied by the distance normalization $((p_i-p_j)/e)$ and by the repulsion factor 0.05 (instead of 0.2).

The result of applying this projection method is showed in Fig. 1. When removal of cells is performed, different meshes are showed unconnectedly. Without any other additional information, this projection method makes possible cluster detection.

Visualization Methods

Using the projection method exposed, traditional Kohonen visualization methods can be implemented using GCS networks with k=2. Each output neuron is painted as a circle in a colour determined by a major parameter. When greyscale is used, normally dark and clear tones are associated with high and low values respectively. The grey scales are relative to the maximum and minimum values taken by the parameter. The nature of the data used to calculate the parameter determines three general types of methods for performing visual analysis of self-organizing maps: distances between synaptic vectors, training patterns projection over the neurons, and individual information about synaptic vectors.

All the experiments have been performed using two groups of simulated data and one group of real multidimensional data (Fig. 2) selected from a scene registered by the ETM+ sensor (Landsat 7). The input signals are defined by the six ETM+ spectral bands with

Figure 1. Output mesh projection during different self-organization process stages of a GCS network trained with bi-dimensional vectors distributed on eleven separate regions.

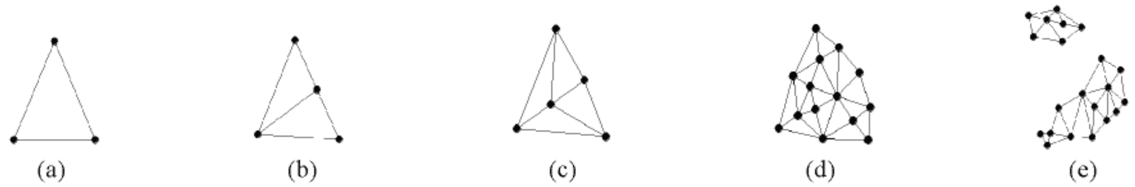
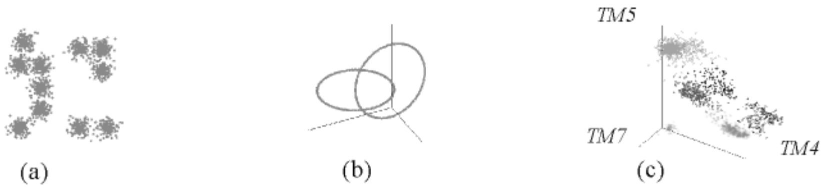


Figure 2. (a) Eleven separate regions in the bi-dimensional plane. (b) Two three dimensional chain-link. (c) Projection of multidimensional data of satellite image.



the same spatial resolution: TM1 to TM5, and TM7. The input data set has a total number of 1800 pixels, 1500 carefully chosen from the original scene and 300 randomly selected. The input vectors are associated to six land cover categories.

Displaying Distances

The adaptation process of GCS networks places the synaptic vectors in regions with high probability density, removing units positioned into regions with a very low probability density. A graphical representation of distances between the synaptic vectors will be a useful tool to detect clusters over the input space. Distance map, unified distance map (U-map), and distance addition map have been implemented to represent distance map information with GCS networks.

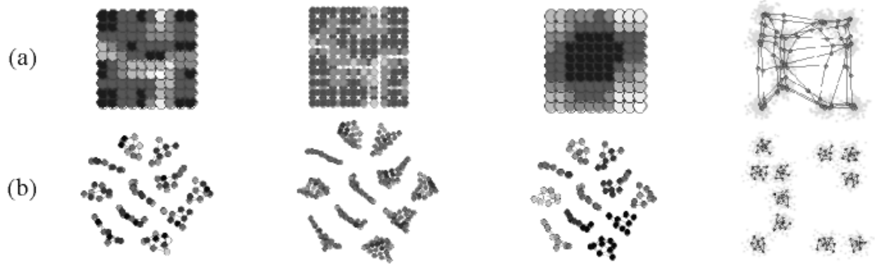
In distance map, the mean distance between the synaptic vector of each neuron and the synaptic vectors of all its direct neighbours is calculated. U-map represents the same information than distance map but, in addition it includes the distance between all the neighbouring neurons (painted in a circle form between each pair of neighbour units). Finally, the sum of the distance between the synaptic vector of a neuron and the synaptic vectors of the rest of units is calculated, when distance addition map is generated. In distance map and U-map, dark zones represent clusters and clear zones boundaries along with them. In distance addition map, neurons with near synaptic vectors appear with similar colour, and boundaries can be detected analyzing the regions where a considerable colour variation exists. Using GCS networks, separated meshes represent different input clusters, usually. Fig. 3 shows an example of these three graphs, compared with the traditional Kohonen's maps, when an eleven separate regions distribution data set is used. GCS network represents eleven clusters in the three graphs, clearly. Distance map and U-map in Kohonen's network show the eleven clusters too, but in distance addition map it is not possible to distinguish them.

Displaying Projections

This technique takes into account the input distribution patterns to generate different values to assign to each neuron. For GCS networks, data histograms and quantization error maps have been implemented.

Generating the histogram, the number of training patterns associated to each neuron is obtained. However, when quantization error graph has to be produced, the sum of the distances between the synaptic vector of a neuron and the input vectors that lies in its *Voronoi* region is calculated. In both graphs, dark and clear zones correspond with high and low probability density areas, respectively, so it can be used in cluster analysis. Fig. 4 shows an example of these two methods compared with those obtained using Kohonen's model when chain-link distribution data set is used. Using Kohonen's model is difficult to distinguish the number of clusters present in the input space. On the other hand, GCS model has generated three output meshes, two of them representing one ring.

Figure 3. From left to right: distance map, U-map (unified distance map), and distance addition map when an eleven separate regions distribution data set is used. (a) Kohonen feature map with 10x10 grid of neurons. (b) GCS network with 100 output neurons. The right column shows the input data and the network projection using the two component values of the synaptic vectors.



Displaying Components

The displaying components technique analyzes each synaptic vector or reference vector component in an individual manner. This kind of graphs offers a visual analysis of the topology preserving of the network, and a possible detection of correlations and dependences between training data components. Direct visualization of synaptic vectors and component planes graphs have been implemented for GCS networks.

Direct visualization map represents each neuron in a circle form within its synaptic vector inside in a graphical manner. This graph can be complemented with anyone of described in the previous sections, enriching its interpretation. A component plane map visualizes an individual component of all the synaptic vectors.

When all the component planes are generated, relations between weights can be appreciated if similar structures appear in identical places of two different component planes. Fig. 5 shows an example of these two displaying methods when multi-band data of satellite image is used. The direct visualization map shows the similarity between neighbouring units synaptic vectors, and, it is interesting distinguish the fact that all the neurons in a cluster have similar synaptic shapes. Furthermore, the integrated information about the distance addition map shows that there is no significant colour variation inside the same cluster. The six component

Figure 4. From left to right: Unified distance map, data histograms and quantization error maps when chain-link distribution data set is used. (a) Kohonen feature map with 10x10 grid of neurons. (b) GCS network with 100 output neurons. The right column shows the input data and the network projection using the three component values of the synaptic vectors.

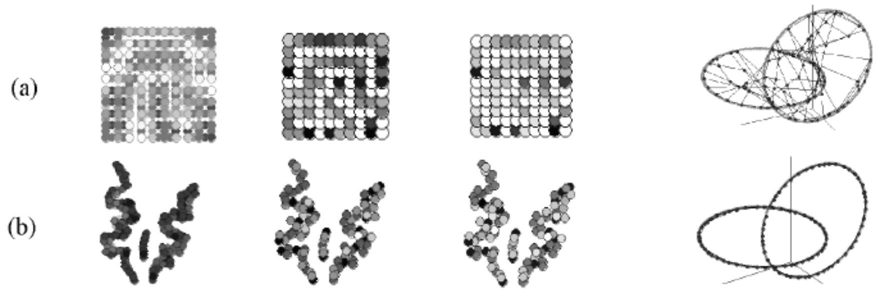
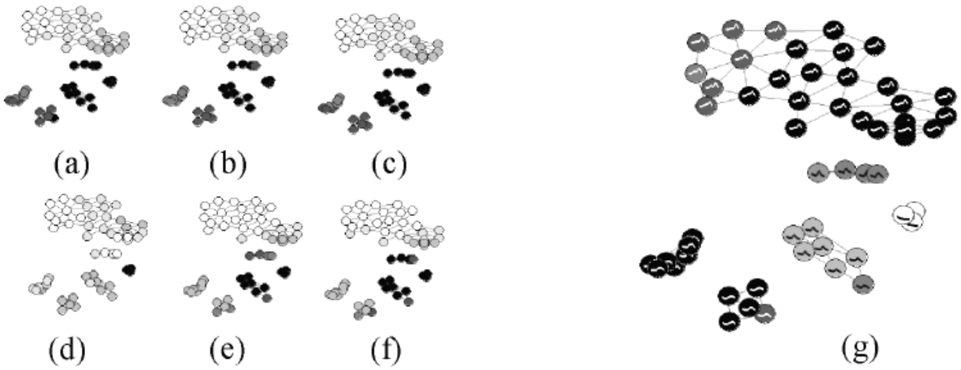


Figure 5. GCS network trained with multidimensional data of satellite image, 54 output neurons. Graphs from (a) to (f) show the component planes for the six elements of the synaptic vectors. (g) Direct visualization map using distance addition map additional information.



plane graphs exhibit possible dependences involving TM1, TM2 and TM3 input vector components and, TM5 and TM7 components too.

Results

Several Kohonen and GCS networks have been trained in order to evaluate and compare the resulting visualization graphs. For the sake of space only a few of these maps have been included here. Fig. 3 and Fig. 4 compare Kohonen and GCS visualizations using distance map, U-map, distance addition map, data histograms and quantization error map. It can be observed that GCS model offers much better graphical results in clusters analysis than Kohonen networks. The removal of units and connections inside low probability distribution areas causes that GCS network presents within a particular cluster the same quality of information that Kohonen network in relation to the entire map. Since it has already been mentioned, the grey scale used in all the maps is relative to the maximum and minimum values taken by the studied parameter. In all the cases the range of values taken by the calculated factor using GCS is minor than using Kohonen maps.

The exposed visualization methods applied to the visual analysis of multidimensional satellite data has given very satisfactory results (Fig 5). All trained GCS networks have been able to generate six sub maps in the output layer (in some case they have arrived up to eight) that identify the six land cover classes present in the sample of data. The direct visualization map and the component plane graphs have demonstrated to be a useful tool for the extraction of knowledge of the multisensorial data.

FUTURE TRENDS

The proposed knowledge visualization method based on GCS networks has results a useful tool for multidimensional data analysis. In order to evaluate the quality of the trained networks we consider necessary to develop some measure techniques (qualitative and quantitative in numerical and graphical format) to analyze the topology preservation obtained. In this way we will be able to validate the information visualized by the methods presented in this paper.

Also it would be interesting to validate these methods of visualisation with new data sets of very high dimensional nature. We need to study the viability of cluster analysis with this projection technique when this class of data samples is used.

CONCLUSION

The exposed embedding method allows multidimensional data to be displayed as two-dimensional grey images. The visual-spatial abilities of human observers can explore these graphical maps to extract interrelations and characteristics in the dataset.

In GCS model the networks size does not have to be specified in advance. During the training process, the size of the network grows and decreases adapting its architecture to the particular characteristics of the training dataset.

Although in GCS networks it is necessary to determine a great number of training factors than in Kohonen model, using the learning modified model the tuning of the training factors values is simplified. In fact, several experiments have been made on datasets of diverse nature using the same values for all the training factors and giving excellent results in all the cases.

Especially notable is the cluster detection during the self-organization process without any other additional information.

REFERENCES

Delgado S., Gonzalo C., Martínez E., & Arquero A. (2004). Improvement of Self-Organizing Maps with Growing Capability for Goodness Evaluation of Multispectral Training Patterns. *IEEE International Proceedings of the Geoscience and Remote Sensing Symposium*. 1, 564-567.

Franzmeier M., Witkowski U., & Rückert U. (2005). Explorative data analysis based on self-organizing maps and automatic map analysis. *Lecture Notes in Computer Science*. 3512, 725-733.

Fritzke, B (1994). Growing Cell Structures – A self-organizing Network for Unsupervised and Supervised Learning. *Neural Networks. 7*(9), 1441-1460.

Fritzke, B (1995). A growing neural gas network learns topologies. *Advances in neural information processing systems. 7*, 625-632.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*. (43), 59-69.

Kohonen, T. (2001). *Self-Organizing Map (*3rd ed). Springer, Berlin Heidelberg New York.

Kraaijveld MA., Mao J., & Jain AK. (1995). A non linear projection method based on Kohonen's topology preserving maps. *IEEE Transactions on Neural Networks. 6*(3), 548-559.

Merlk D., & Rauber A. (1997). Alternative ways for cluster visualization in self-organizing maps. *Workshop on Self-Organizing Maps*, Helsinki, Finland.

Rossi, F. (2006). Visual data mining and machine learning. *Proceedings of the European Symposium on Artificial Neural Networks*, Bruges, Belgium. 251-264.

Rubio M., & Giménez V. (2003). New methods for self-organizing map visual analysis. *Neural Computation & Applications*. 12, 142-152.

Tukey, JW. (1977). *Exploratory data analysis*. Addison-Wesley, Reading, MA.

Ultsch A., & Siemon HP. (1990). Kohonen self-organizing feature maps for exploratory data analysis. *Proceedings of the International Neural Network*, Dordrecht, The Nederlands.

Vesanto, J. (1999). SOM-based visualization methods. *Intelligent Data Analysis*, Elsevier Science, 3(2), 111-126.

Walker AJ., Cross SS., & Harrison RF. (1999). Visualisation of biomedical datasets by use of growing cell structure networks: a novel diagnostic classification technique. *The Lancer*, Academic Research Library. 1518-1521.

KEY TERMS

Artificial Neural Networks: An interconnected group of units or neurons that uses a mathematical model for information processing based on a connectionist approach to computation.

Data Mining: The application of analytical methods and tools to data for the purpose of identifying patterns, relationships or obtaining systems that perform useful tasks such as classification, prediction, estimation, or affinity grouping.

Exploratory Data Analysis: Philosophy about how a data analysis should be carried out. Exploratory data analysis employs a variety of techniques (mostly graphical) to extract the knowledge inherent to the data.

Growing Cell Structures: Growing variant of the self-organizing map model, with the peculiarity of dynamically adapts the size and connections of the output layer to the characteristics of the training patterns.

Knowledge Visualization: The creation and communication of knowledge through the use of computer and non-computer-based, complementary, graphic representation techniques.

Self-Organizing Map: A subtype of artificial neural network. It is trained using unsupervised learning to produce low dimensional representation of the training samples while preserving the topological properties of the input space

Unsupervised Learning: Method of machine learning where a model is fit to observations. It is distinguished from supervised learning by the fact that there is no a priori output.