Evaluating Growing Self-Organizing Maps for Satellite Image Time Series Clustering

Rodrigo S. S. Adeu[1], **Karine R. Ferreira**[1], **Pedro R. Andrade**[1], **Lorena Santos**[1]

[1] National Institute for Space Research (INPE),
Astronautas Avenue 1758, 12227-010,
São José dos Campos – São Paulo – Brazil

rodrigo.sales@embraer.com.br,
{karine.ferreira,pedro.andrade,lorena.santos}@inpe.br

Abstract. *In recent years, analysis of time series extracted from Earth observation satellite images has been widely used to produce land use and cover information. In time series analysis, clustering is a common technique performed to discover patterns on data sets. Self-Organizing Maps (SOM) neural network is a suitable method for such task. However, a critical limitation of SOM is that its map structure size must be predetermined. This limitation has been addressed by Growing SOM method. This paper presents an ongoing work on evaluating Growing SOM for Earth observation satellite image time series clustering.*

1. Introduction

Machine learning methods, such as Support Vector Machine (SVM) and Random Forest (RF), have been used to classify Earth observation image time series in order to produce land use and cover change maps [Picoli et al. 2018]. Most of these methods are based on supervised machine learning algorithms that require a training phase using labelled land use and cover samples. Selecting representative samples is crucial to obtain good accuracy in the classifications.

To better select land use and cover samples from satellite image time series, Santos et al. [Santos et al. 2019] propose a method based on the Self-Organizing Map (SOM) neural network [Kohonen et al. 2001]. The method uses SOM in the training phase to estimate the quality of the land use and cover samples as well as to evaluate which spectral bands and vegetation indexes are best suitable to differentiate land use and cover classes. This method explores two main features of SOM: (1) the capacity of mapping a high-dimensional input space into a two-dimensional grid; and (2) the topological preservation of neighborhood, which generates spatial clusters of similar patterns in the output space.

Despite its advantages, SOM has a characteristic that limits its potential. It uses a fixed network architecture in terms of number and arrangement of neural processing elements which have to be predefined. The need to predetermine the size of the network is not considered a simple task. Simulations have to be run several times on different network sizes to find an appropriate network structure [Flexer 2001, Kohonen et al. 2001].

This paper presents an ongoing work to evaluate Growing SOM (GSOM) as an alternative to traditional SOM for satellite image time series clustering. GSOM method was originally proposed to address the SOM limitation on predeterminng the map size [Alahakoon et al. 2000]. This work aims to contribute to land use and cover research area by validating an approach that avoids this additional parameter.

2. Satellite image time series clustering using SOM

Since remote sensing satellites constantly revisit the same place, it is possible to calibrate images so that measures of the same place at different times are comparable (Figure 1(a)). Such images can be organized to compose a three-dimensional array in space-time. From a data analysis perspective, each pixel location (x, y) at consecutive times $t_1, ..., t_m$ makes up a satellite image time series, such as the one in Figure 1(b). From these time series, we can extract land use and land cover information.

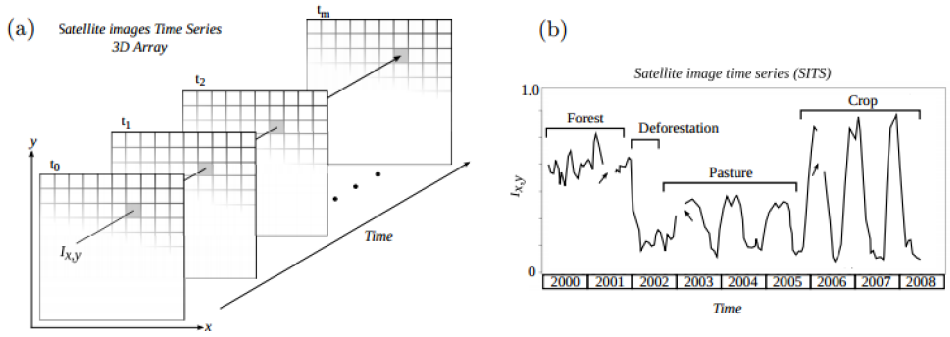

Figure 1. Deriving time series from Earth observation satellite images: (a) A dimensional array of satellite images, (b) vegetation index time series at a fixed (x,y) pixel location [Maus et al. 2016].

Clustering is a common technique performed to discover intrinsic patterns on time series data sets, by grouping similar time series together based on a certain similarity measure. SOM has been widely used for time series clustering in various domains, such as meteorology and oceanography [Liao 2005, Mwasiagi 2011, Liu et al. 2016, Pearce et al. 2014].

Santos et al. [Santos et al. 2019] propose a methodology that can be used as an exploratory analysis tool for land use and cover samples from remote sensing image time series using SOM. This methodology provides means to detect sample outliers using neighborhood analysis. For example, Figure 2 shows neurons labelled as Soy-Corn, Millet-Cotton, and Soy-Sunflower in the middle of a region classified as Soy-Cotton. Since the classes of the samples in such neurons match the input classes, two possibilities can be considered. The first one is that these input samples are outliers, possibly by an error in the classification of the samples. A second hypothesis is that the samples of different classes are so similar that such classes cannot separated by SOM using the current input samples and attributes.

SOM uses a fixed neuron map whose size must be predefined. The need to predetermine the size of the network is not considered a simple task. The literature shows that determining grid size for SOM is an empirical process [Flexer 2001, Kohonen et al. 2001]. Simulations have to be run several times on different network sizes to find an appropriate network structure. On the next section, alternatives to dynamically evolve the SOM grid size are presented.

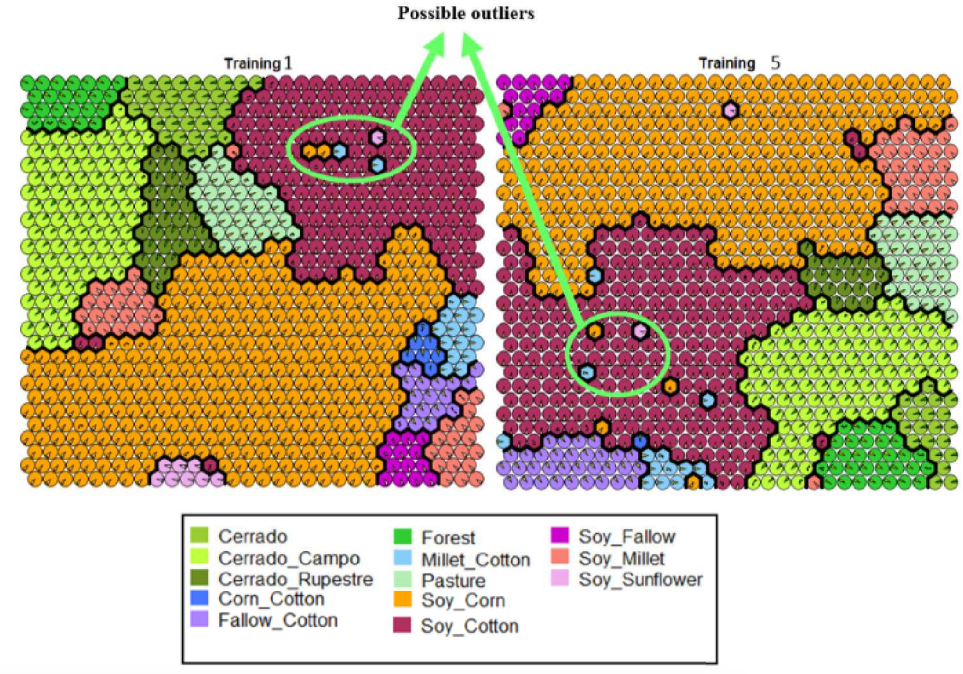

Figure 2. Evaluating land use and cover samples. Final SOM clustering and possible outliers [Santos et al. 2019].

3. Growing SOM method and implementations

The Growing Self-Organizing Map (GSOM) is a neural network with a dynamic structure designed to solve the limitation of predetermined network size in conventional SOMs. The main difference between the two methods is that SOM attempts to fit a data set into a predefined structure by self-organizing its node weights as well as possible within its fixed borders. In GSOM, the borders of the network are expandable. It might generate nodes whenever needed, expanding the network outwards [Alahakoon et al. 2000]. GSOM is parameterized by a Spread Factor, which is independent of the dimensionality of the data. It can be used as a controlling measure for generating maps with different dimensionality, which can in turn be compared and analyzed with better accuracy.

GSOM achieves the same amount of spread that traditional SOM, with a lesser number of nodes, providing a useful advantage in mapping large data sets. In addition, such flexible structure provides a better visualization of the groups in the data and attracts attention to outliers by branching them out. It also preserves the neighborhood while growing the map. GSOM keeps the simplicity and facility of SOM use, expanding its usefulness by dynamically generating the map structure [Alahakoon et al. 2000]. These characteristics, combined with the detailed specification presented in [Alahakoon et al. 2000], make GSOM a feasible alternative for SOM.

As a starting point, we evaluated three available implementations of the GSOM algorithms: PyGSOM Python package [Ludwig 2016], GSOM Python package [Mendis 2015] and GrowingSOM R package [Hunziker 2018]. These implementations

were tested, but the results were not satisfactory. In some cases the performance were not acceptable, or the algorithm specification was not precisely implemented.

In the PyGSOM Python package, different possible approaches for the GSOM algorithm were used in the implementation, resulting on a mixed solution. Ludwig states that this implementation should not be taken as a reference [Ludwig 2016].

The GSOM Python package stores only the last sample associated to each neuron, instead of all the samples. As a consequence, visualization of the best matching units were based on the last classified sample, not on the most common. Furthermore, after running examples, this solution seems to not respect the neighborhood while growing the grid.

During the GrowingSOM R package testing, we have noticed that this implementation does not store the relationship between the samples and the neuron associated to them. The visualization features and the developed public interfaces are also limited. But the main concern of this implementation was the training performance. As stated by [Kane et al. 2013], R has a limitation on processing large objects and is not designed for working with data structures greater than 10% - 20% of a computer RAM memory, resulting in performance issues. As this solution is fully implemented in R, and the main goal for this work is the clustering of satellite image time series, the performance was not acceptable.

4. Proposed solution and preliminary results

As described in section 3, the available GSOM implementations were tested and none of them was working as expected. So, we decided to develop a new R package with the GSOM algorithm proposed by [Alahakoon et al. 2000] using the Kohonen R package [Wehrens and Buydens 2007]. The Kohonen R package is available on CRAN and provides the original SOM functionality with good performance due to its Rcpp implementation [Eddelbuettel and François 2011]. It is recognized as a stable implementation of SOM by the community. It aims to provide simple-to-use functions for SOM, with specific emphasis on visualization.

In this work, a new GSOM R package was developed, upgrading the Kohonen R package implementation by cloning its current code and implementing the GSOM functionalities inside its C++ code. The proposal is take advantage of the already developed SOM benefits, and implement only the differences needed to provide the GSOM capabilities. These modifications are still in progress. Besides that, the overall performance of the algorithm was acceptable, as the time spend on the growing phase added less than 5% in the algorithm execution time, for the related data set. Visualization features provided by the original package could also be used without further adaptations.

Figure 3 presents the result of the new GSOM R package developed in this work using the same sample data set used by Santos et al. [Santos et al. 2019]. In this figure, we can observe the growing grid capabilities and the generalization capability of the neurons. Neuron 29 illustrates a cluster of 15 time series in the same Neuron, most of them belonging to Pasture class. However, Neuron 09 clustered 651 time series of 8 different classes, indicating possible generalization issues on this neuron. Alternative neuron weights initialization has been tested as possible alternative to address this issue.

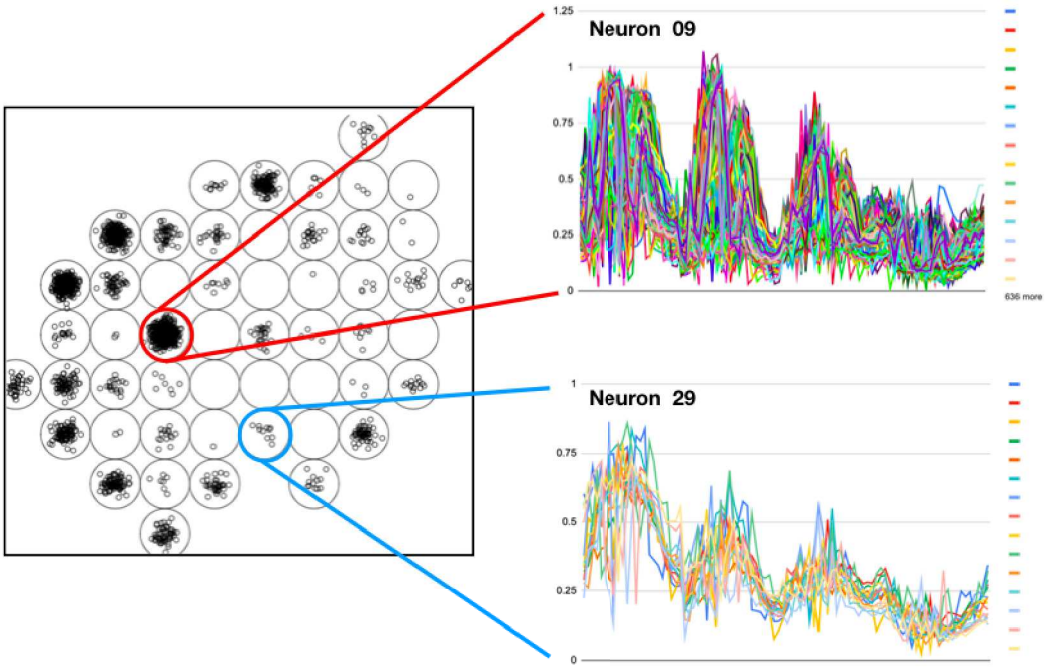

Figure 3. Map and clusters generated by the new GSOM R package.

5. Final remarks

This is an ongoing work and the preliminary results presented in section 4 indicate that the GSOM algorithm is promising for clustering time series extracted from Earth observation satellite images. We can observe in Figure 3 that the grid map grew as expected and many neurons grouped similar time series together.

After finishing the new GSOM R package development, the objective is to check the GSOM improvements by performing experiments using time series extracted from MODIS sensor of the Terra satellite, developed by NASA. The study area of these experiment will be the Mato Grosso state whose samples include three Brazilian biomes: Amazonia, Cerrado, and Pantanal. Several GSOM executions will be runned, comparing the results with the fixed 15 x 15, 40 x 40 and 50 x 50 SOMs obtained by Santos et al. [Santos et al. 2019].

The goal of these experiments will be to check the generated grid size, comparing the number of neurons used by GSOM with the number of neurons used by SOM. Besides that, the sample density on the GSOM neurons will be analyzed, and the accuracy of the GSOM clustering will be measured and compared with the accuracy obtained by SOM.

References

Alahakoon, D., Halgamuge, S. K., and Srinivasan, B. (2000). Dynamic self-organizing maps with controlled growth for knowledge discovery. *IEEE Transactions on Neural Networks*, 11:601–614.

Eddelbuettel, D. and François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18.

Flexer, A. (2001). On the use of self-organizing maps for clustering and visualization. *Journal Intelligent Data Analysis*, 5:373 – 384.

Hunziker, A. (2018). Growingsom r package. available at: https://github.com/alexhunziker/growingsom.

Kane, M. J., Emerson, J. W., and Weston, S. (2013). Scalable strategies for computing with massive data. *Journal of Statistical Software*, pages 1–19.

Kohonen, T., Schroeder, M. R., and Huang, T. S. (2001). *Self-Organizing Maps*. Springer-Verlag, 3rd edition edition.

Liao, T. W. (2005). Clustering of time series data - a survey. *Pattern Recognition*, 38:1857–1874.

Liu, Y., Weisberg, R. H., Vignudelli, S., and Mitchum, G. T. (2016). Patterns of the loop current system and regions of sea surface height variability in the eastern gulf of mexico revealed by the self-organizing maps. *Journal of Geophysical Research: Oceans*, 121(4):2347–2366.

Ludwig, P. (2016). Pygsom - a gsom (growing self-organizing map) implementation for python. available at: https://github.com/philippludwig/pygsom.

Maus, V., Camara, G., Cartaxo, R., Sanchez, A., Ramos, M., and Queiroz, G. (2016). A time-weighted dynamic time warping method for land-use and land-cover mapping. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(8):3729 – 3739.

Mendis, L. (2015). Gsom - the growing self organizing map implementation on python. available at: https://github.com/anantadata/gsom.

Mwasiagi, J. I. (2011). Self organizing maps - applications and novel algorithm design. *InTech*, page 253–72.

Pearce, J. L.and Waller, L. A., Chang, H. H., Klein, M., Mulholland, J. A., Sarnat, J. A., Sarnat, S. E., Strickland, M. J., and Tolbert, P. E. (2014). Using self-organizing maps to develop ambient air quality classifications: a time series example. *Environmental health: a global access science source*, 13:56.

Picoli, M., Camara, G., Sanches, I., Simoes, R., Carvalho, A., Maciel, A., Coutinho, A., Esquerdo, J., Antunes, J., Begotti, R., Arvor, D., and Almeida, C. (2018). Big earth observation time series analysis for monitoring brazilian agriculture. *ISPRS Journal of Photogrammetry and Remote Sensing*, 145:328 – 339.

Santos, L. A., Ferreira, K. R., Picoli, M., and Camara, G. (2019). Self-organizing maps in earth observation data cubes analysis. *International Workshop on Self-Organizing Maps*, pages 70–79.

Wehrens, R. and Buydens, L. (2007). Self and super-organizing maps in r: The kohonen package. *Journal of Statistical Software*, 21.