Computers and Geosciences 119 (2018) 9–17

Contents lists available at ScienceDirect

Computers and Geosciences

journal homepage: www.elsevier.com/locate/cageo

Research paper

Interpreting Self-Organizing Map errors in the classification of ocean patterns

Frano Matić[a,*], Hrvoje Kalinić[b], Ivica Vilibić[a]

[a] Institute of Oceanography and Fisheries, Šetalište I. Meštrovića 63, 21000, Split, Croatia
[b] Faculty of Science, University of Split, Ruđera Boškovića 33, 21000, Split, Croatia

ARTICLE INFO

Keywords:
Self organizing maps
Error estimation
Time series analysis
Outliers and extreme data
Climate regimes
Oceanography
Meteorology

ABSTRACT

The paper aims to introduce quality measures that can evaluate how well the Self-organizing Maps method performs in transitional stages. The errors have been computed with respect to the spatial and temporal properties of the data and in relation to the data gap significance. Temperature and salinity data collected in the central Adriatic Sea at six stations during 196 field cruises carried out between 1963 and 2011 have been used for the mapping of ocean patterns and computation of the respective errors. The errors resemble both the stability of ocean regimes and variability of patterns that are documented in the investigated region. As the data collection methodology and approach have changed over time, the errors may be a good indication for the presence of bad data in a series, which may then be controlled by other quality-check techniques.

1. Introduction

Machine learning represents an optimization process in which a solution (optimal or extremal) has been searched. Supervised machine learning is a well-defined process; by definition, there exists a solution to the problem. Unsupervised machine learning has no “golden rule”, and solutions are not labelled; thus, there is no predefined error metric to evaluate the quality of learning, making unsupervised learning a much less well-defined problem (Murphy, 2012). Therefore, unsupervised learning has a lack of the objective evaluation of the accuracy (Hastie et al., 2009; Murphy, 2012). Hastie et al. (2009) state that with unsupervised learning, we have no measurements of the outcome and there is no direct measure of learning success, pointing out that the main task of unsupervised learning is to organize (cluster) the data. However, Duda et al. (2001) emphasize that the problem of unsupervised learning is too important to abandon just because exact solutions are hard to find. Moreover, the lack of exact solutions or objective measures of an evaluation comes as a side effect of unsupervised learning, which is often implemented on problems with large amounts of unlabelled data (Duda et al., 2001), i.e., on problems that do not utilize a gold standard and/or when there is no ground truthing available (e.g., Bruzzone and Prieto, 2000; Volkovs and Zemel, 2014; Holmes, 2014).

The Self-Organizing Maps (SOM) method, like principal component analysis (PCA), can be utilized as a dimensionality reduction technique. The reliability assessment of PCA dimensionality reduction is usually referred by unexplained variance (variance unexplained by the proposed model in the reduced space), and it is given as the sum of squared distances (Preisendorfer and Mobley, 1988). In contrast to PCA, SOM is a mapping technique that performs nonlinear dimensionality reduction by mapping the input data to the nearest (the winning) neuron in the SOM lattice. Whenever an interval (or in an N dimensional space N-D sphere) is represented as one point (the winning neuron), a discretization or quantization is done since the method maps a larger (continuous) set to a smaller (discrete) set. Thus, the quantization error, which is also usually measured as the sum of squared differences (SSD), denotes the distortion (Vesanto et al., 2003).

In real case applications, and specifically in geosciences, SOM is used to identify patterns and regimes of the system (atmosphere, ocean, etc.), which are represented by the neurons in the SOM lattice (Kirk and Zurada, 2004; Bação et al., 2005; Gorricha and Lobo, 2012; Kalinić et al., 2015a). As there is no any statistical a priori request for the input data for the successful application of an SOM model, the SOM model also became widely used in marine biology (Bandelj et al., 2008; Šolić et al., 2018) and environmental studies (Ley et al., 2011; Toth, 2013). The SOM model was also used to overcome insufficient or multiple information that can be extracted by either classical statistical methods or Monte-Carlo simulations (Herbst et al., 2009).

Reliability of SOM can be utilized by analyzing quality of convergence and generated maps (Tatoian and Hamel, 2018), analyzing different quantization and topographic errors (Uriarte and Martín, 2008) and quality of topology preservation for different initial

* Corresponding author.
E-mail address: fmatic@izor.hr (F. Matić).

https://doi.org/10.1016/j.cageo.2018.06.006
Received 12 January 2018; Received in revised form 11 June 2018; Accepted 14 June 2018
Available online 18 June 2018
0098-3004/

condition (Villmann et al., 1997), map size and sampling variability (Cottrell et al., 2001).

The widely used method to get the quality of the SOM model is trying different SOM architectures, mainly changing the size of the reduced space and choosing the one that maximizes the information content (Liu et al., 2006; Vilibić et al., 2011). A bootstrap method may be used for choosing the best matrix size for a certain dataset (Cottrell et al., 2001; Bodt de et al., 2002). Additionally, it is possible to apply a clustering method over SOM with a large number of map units and then validate the goodness of clusterization. Finally, a double-step cluster procedure can indirectly give the quality of SOM (Vesanto and Alhoniemi, 2000; Solidoro et al., 2007, 2009; Šolić et al., 2018).

However, research studies were mainly focused on the determination of patterns and their sensitivity/reliability (Joutsiniemi et al., 1995; Matsuda et al., 2014; Kalinić et al., 2015b; Vilibić et al., 2016a, 2016b), while the question of how accurately patterns represent rare and odd events that usually occur in transitional stages of the system were slightly out of focus. Normally, the goodness of fit of the SOM method to the data is assessed through quantization error and topographic error (Kohonen, 2001; Liu et al., 2006), of which the first is the average distance between each data vector and the winning neuron, while the second is the percentage of the data vectors for which the winning neuron and its runner-up are not neighbouring units. The lower are both measures the better is quality of the SOM mapping. However, none of these measures is equivalent to the common measures used in quality assessment of data mapping in geosciences, where bias and root-mean-square-error is normally used for assessment of the goodness of fit of a model to the data (Hyndman and Koehler, 2006). In this paper, we aim to provide a measure that is equivalent to the classical statistical measures and can evaluate how well the SOM method performs in transitional stages/oscillations, i.e., how accurately the SOM patterns represent the real state of the system. We named the measure as the winning error, and it measures between the winning neuron and the respective data. We tested the methodology on the long-term oceanographic series collected in the Adriatic Sea over several decades (1963–2011) and focused on the variability and distribution of error estimates. Section 2 describes the methodology and the data used in the study, Section 3 presents the results and discussion, Section 4 presents the conclusions.

2. Data and methods

2.1. SOM method, performance metrics

Traditionally, Self-Organizing Maps are classified as artificial neural networks that differ from typical artificial neural networks in the sense that they utilize competitive algorithms rather than error-correction. In general, SOM is an unsupervised learning method that can be related to k-means clustering but preserves the topology of input data, which makes it particularly appealing as a dimensionality reduction technique, especially given that SOM does not require a priori assumptions on the nature of data. Thanks to these properties, SOM was presented as a method that reduces dimensionality using the nonlinear projection of input data onto a two-dimensional grid. The input data for the SOM model is matrix D_{MxN} with vectors containing ordered information, data, from one experiment. In the learning stage of the SOM process, the data is introduced to SOM lattice with k neurons and the winning neurons ($C_1 \ldots C_k$) are updated to match the input data. For this reason the winning neurons are often referred to as best matching units (BMUs) or codebook vectors. In the mapping phase, the data, given by the column vector (D_i), is associated to the nearest among the winning neurons:

$$K_k = \{D_i | \sqrt{(D_i - C_k)^2} \leq \sqrt{(D_i - C_l)^2}, \quad \forall\, l \in [1, k]\,\}. \tag{1}$$

Each column vector (D_i) describes the measured state of the system at time i, denoted as S_i. The input data matrix (D) can now be represented in compressed form if neurons (C_i), are used to represent the system states. After analysis, one neuron (C_k) with the shortest Euclidian distance is associated with the data vector (D_i), e. q. quantization is performed, and each state of the system is substituted by its approximation:

$$\hat{S} = \arg\min_{C_k} \sqrt{(C_k - D_i)^2}\,. \tag{2}$$

The distance varies between the data and shows the quality of the data vector projection to the winning neuron. The quality of this nonlinear mapping is normally assessed by two measures: the quantization error that quantifies the average distance between each data vector and the BMU, and the topographic error that gives the percentage of the data vectors for which the first BMU and the second BMU are not neighbouring units. The lower the errors are, the better the SOM model fits the input data (Kohonen, 1982, 2001, 2013).

However, these measures are somehow not comparable to measures that are frequently used in geosciences, such as root-mean-square-error, which is normally used for assessing variability in the data coming from linear systems. Consequently, a large number of applications of the SOM to geosciences and environmental data tend to avoid an assessment of representativeness of the SOM method to the data. Therefore, to assess the variability of the data, we introduced the codebook error (CBE) as a normalized quantization error of a specific neuron, i.e., CBE is given as the error between each codebook vector (or BMU) and the elements of the set:

$$CBE_k = \sqrt{\frac{1}{\#K_k} \sum_{D_i \varepsilon K_k} (D_i - C_k).*(D_i - C_k)}\,, \tag{3}$$

where .* stands for the Hadamard (elementwise) product of vectors, and $\#K_k$ stands for the cardinality (the number of elements) of the set K_k. The normalization is done with respect to the number of elements of the set, i.e., as the average quantization error for a given codebook vector or set representative. In contrast to quantization error, which is equal to the overall distance between the winning neurons and the data, the CBE is calculated as the root-mean-square-error between one codebook vector and data in time when that codebook vector is the winning neuron.

Thus, CBE_k explains the expected quantization error for each state of the system and can be used to visualize the spatial distribution of RMSE if the winning neuron is used to represent the data (Kalinić et al., 2015b). When the winning neuron (BMU) has been used to represent the data, a quantization has been performed. Thus, CBE can be observed as a special method to measure the quantization error or distortion made when data is represented by its associated BMUs. While CBE well explains the spatial distribution of the error, it might not well explain the temporal changes that occur when the system is in a transitional stage. The transitional stage can vary in duration, have oscillations and overall explain different events. While the temporal change of the system can be visualized by the temporal evolution of BMUs, neither quantization error nor CBE answer if the temporal evolution of BMUs reliably represents the system in its transitional stages.

To measure how the SOM method performs at a specific time interval i, we define the winning element error (WEE):

$$WEE_i = \sqrt{(D_i - \hat{S}_i).*(D_i - \hat{S}_i)}\,. \tag{4}$$

Notice that WEE is a vector. To assess the error in transitional stages, scalar quantization error could be measured at each time interval:

$$\varepsilon_i = \|S_i - \hat{S}_i\|_2 = \|D_i - \hat{S}_i\|_2. \tag{5}$$

which is equivalent to: