Information Sciences 364–365 (2016) 213–221

Contents lists available at ScienceDirect

Information Sciences

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

SOM: Stochastic initialization versus principal components

Ayodeji A. Akinduko, Evgeny M. Mirkes, Alexander N. Gorban*

University of Leicester, Leicester, UK

ARTICLE INFO

Article history:
Received 10 April 2015
Revised 18 September 2015
Accepted 6 October 2015
Available online 27 October 2015

Keywords:
SOM
PCA
Random initiation
Benchmark test
Manifold learning

ABSTRACT

Selection of a good initial approximation is a well known problem for all iterative methods of data approximation, from *k*-means to Self-Organizing Maps (SOM) and manifold learning. The quality of the resulting data approximation depends on the initial approximation. Principal components are popular as an initial approximation for many methods of nonlinear dimensionality reduction because its convenience and exact reproducibility of the results. Nevertheless, the reports about the results of the principal component initialization are controversial.

In this work, we separate datasets into two classes: *quasilinear* and *essentially nonlinear* datasets. We demonstrate on learning of one-dimensional SOM (models of principal curves) that for the quasilinear datasets the principal component initialization of the self-organizing maps is systematically better than the random initialization, whereas for the essentially nonlinear datasets the random initialization may perform better. Performance is evaluated by the fraction of variance unexplained in numerical experiments.

1. Introduction

Principal components produce the best linear approximations of datasets ("lines and planes of closest fit to systems of points" [24]). These lines and planes are popular as initial approximations for many methods of nonlinear dimensionality reduction [13,17,19] because their convenience and exact reproducibility of the results. The quality of the resulting data approximation depends on the initial approximation but the systematic analysis of this dependence requires usually too much efforts and the reports are often controversial.

In this work, we analyze initialization of Self Organized Maps (SOM). We test and systematically compare two main approaches to initial approximation of SOM. Originally, Kohonen [18] has proposed random initiation of SOM weights but recently the principal component initialization (PCI), in which the initial map weights are chosen from the space of the first principal components, has become rather popular [5]. Nevertheless, some authors have criticized PCI [4,29] (see also discussion of PCI in recent work [30]). For example, the initialization procedure is expected to perform much better if there are more nodes in the areas where dense clusters are expected and less nodes in empty areas. In practical applications, SOM initialization is often performed in several different ways [25].

In this paper, the performance of random initialization (RI) approach is compared to that of PCI for one-dimensional SOM (models of principal curves). Performance is evaluated by the Fraction of Variance Unexplained (FVU). Datasets were classified into linear, quasilinear and nonlinear [14,15]. It was observed that RI systematically performs better for nonlinear datasets; however the performance of PCI approach remains inconclusive for quasilinear datasets.

SOM can be considered as a nonlinear generalization of the principal component analysis [32,33]. Originally developed by Kohonen [18] for visualization of distribution of metric vectors, SOM found many applications in data exploration, especially in

* Corresponding author. Tel.: +441162231433.
E-mail addresses: aaa78@le.ac.uk (A.A. Akinduko), em322@le.ac.uk (E.M. Mirkes), ag153@le.ac.uk, gorbanster@gmail.com (A.N. Gorban).

http://dx.doi.org/10.1016/j.ins.2015.10.013
0020-0255/

data visualization, vector quantization and dimension reduction. However, like for clustering algorithms [12,26], the quality of learning of SOM is greatly influenced by the initial conditions: initial weight of the map, the neighborhood function, the learning rate, sequence of training vector and the number of iterations [18,28]. Several initialization approaches have been developed and can be broadly grouped into two classes: the random initialization and the data analysis based initialization [4]. Due to many possible initial configurations when using random approach, several attempts are usually made and the best initial configuration is adopted.

For the data analysis based approach, certain statistical data analysis and data classification methods are used to determine the initial configuration; a popular method is selecting the initial weights from the space spanned by the linear principal component. Modification to the PCA approach was done [4] and over the years other initialization methods have been proposed. An example is given by Fort et al. [11]. Careful testing is needed for comparison of different SOM initialization strategies.

In this paper, we consider the performance in terms of the quality of learning of SOM using the Random Initialization (RI) method (in which the initial weights are randomly selected from the sample data) and the Principal Component Initialization (PCI) method. The quality of learning is determined by the fraction of variance unexplained [22]. To ensure an exhaustive study, synthetic data sets distributed along various shapes of dimension two are considered in this study and the map is one-dimensional (1D). 1D SOMs are important, for example, for approximation of principal curves. The experiment was performed using the PCA, SOM and Growing SOM (GSOM) applet available online [22] and can be reproduced. The SOM learning has been done with the same neighborhood function and learning rate for both initialization approaches. Therefore, the two methods are subject to the same conditions which could influence the learning outcome of our study. To marginalize the effect of the sequence of training vectors, the applet adopts the batch learning SOM algorithm [10,11,18] described in the next Section. We also test our findings on several popular multidimensional benchmarks and on two-dimensional (2D) SOM.

For the random initialization approach, the space of initial starting weights was sampled; this is because as the size of the data set n increases, the possible choice of initial configuration for a given number of nodes k becomes enormous (n^k). The PCI was done using regular grid on the first principal component with equal variance (Mirkes, 2011). For each data set and initialization approach, the data set was trained using three or four different values of k. We use a heuristic classification of datasets in three classes, linear, quasilinear and essentially nonlinear [14,15], to organize the case study and to represent the results. We describe below the used versions of the SOM algorithms in detail in order to provide the *reproducibility* of the case study.

It is demonstrated that for essentially nonlinear patterns the widely accepted presumption about advantages of PCI SOM initialization is not universal. RI (possibly with several reinitialization) often performs better than PCI.

2. Background

2.1. SOM algorithm

SOM is an artificial neural network which has a feed-forward structure with a single computational layer. Each neuron in the map is connected to all the input nodes. The classical on-line SOM algorithm can be summarised as follows:

1. Initialization: Initial weights are assigned to all the connection $w_j(0)$.
2. Competition: all nodes compete for the ownership of the input pattern. Using the Euclidean distance as criterion, the neuron with the minimum-distance wins.

$$j^* = \arg\min_{1\le j\le k} \|x(t) - w_j(t)\|,$$

   where $x(t)$ is the input pattern at time t, $w_j(t)$ is jth coding vector at time t, k is the number of nodes.
3. Cooperation: the winning neuron also excites its neighboring neurons (topologically close neurons). The closeness of the ith and jth neurons is measured by the neighborhood function $\eta_{ji}(t)$: $\eta_{ii} = 1$, $\eta_{ji} \to 0$ for large $|i - j|$.
4. Learning Process (Adaptation): The winning neuron and the neighbors are adjusted with the rule given below:

$$w_i(t+1) = w_i(t) + \alpha(t)\eta_{j^*i}(x(t) - w_i(t)),$$

   Thus, the weight of the winning neuron and its neighbors are adjusted towards the input patterns however the neighbors have their weights adjusted with a value less than the winning neuron. This action helps to preserve the topology of the map. A scalar factor $\alpha(t)$ (the speed of learning) defines the size of the correction; for most realizations, its value decreases with time t [18].

2.2. The batch algorithm

We use the batch algorithm of the SOM learning. This is a version of the SOM algorithm in which the whole training set is presented to the map before the weights are adjusted with the net effect over the samples [10,18,21]. The algorithm is given below (after the standard initialization).

1. Put the set of data point associated with each node equal to empty set: $C_i = \emptyset$.