Clustering-Based Adaptive Self-Organizing Map

Dominik Olszewski(✉)

Faculty of Electrical Engineering, Warsaw University of Technology, Warsaw, Poland
dominik.olszewski@ee.pw.edu.pl

Abstract. We propose an improvement of the Self-Organizing Map (SOM). In our version of SOM, the neighborhood widths of the Best Matching Units (BMUs) are computed on the basis of the data density and scattering in the input data space. The density and scattering are expressed by the values of the inner-cluster variances, which are obtained after the preliminary input data clustering. The experiments conducted on the two real datasets evaluated the proposed approach on the basis of a comparison with the three reference data visualization methods. By reporting the superiority of our technique over the other tested algorithms, we confirmed the effectiveness and accuracy of the introduced solution.

Keywords: Self-Organizing Map · Adaptive self-organizing map · clustering · Neighborhood width · Gaussian kernel · Visualization

1 Introduction

The Self-Organizing Map (SOM) [3] is a type of an artificial neural network architecture, however, at the same time, it may be recognized as a data visualization technique. The term data visualization refers in our research to a linear or non-linear projection from an original input high-dimensional space onto a resulting output 2- or 3-dimensional data space. Consequently, any data visualization formulated in the following way can be treated as a particular case of a dimensionality reduction problem, where the output number of dimensions is 2 or 3, typically 2. The SOM technique generates a 2-dimensional map structure. The location of points in 2-dimensional grid aims to reflect the similarities between the corresponding samples in an input multidimensional space. Therefore, the SOM algorithm allows for visualization of relationships between samples in multidimensional space.

1.1 Our Proposal

We propose an adaptive rule for the determination of the neighborhood widths of the SOM's BMUs during the SOM's training. The rule is formulated on the basis

© Springer Nature Switzerland AG 2021
L. Rutkowski et al. (Eds.): ICAISC 2021, LNAI 12854, pp. 182–192, 2021.
https://doi.org/10.1007/978-3-030-87986-0_16

of the preliminary data clustering in the input space. After forming of the input data clusters, the inner-cluster variances for each of the generated clusters are calculated and utilized afterward as a basis for the SOM's BMUs' neighborhood widths computation.

In our method, the neighborhood widths are determined independently for each BMU neuron in the SOM grid. The neighborhood of BMU is mathematically described using the Gaussian kernel function, where the radius of the Gaussian kernel is calculated as a result of initial data clustering in the input space. This is achieved in this way that the inner-cluster variances for each of the input data clusters are utilized as the basis for the radius of the Gaussian kernel.

2 Related Work

The SOM visualization technique has been extensively studied, and numerous improvements and extensions have been developed, including the Growing Hierarchical SOM (GHSOM) [12], the asymmetric SOM [4,7,10], and the adaptive SOM [1,5,9,11,13], to name a few. Naturally, the adaptive SOM versions are of particular interest for the purposes of our research.

An approach allowing to gain a control over the neurons' neighborhood widths in SOM delivered in the paper [13] is the magnification control approach. The issue is thoroughly studied by the authors of [13], where the three learning rule modifications for SOM are considered, namely, the localized learning, the winner-relaxing learning, and the concave-convex learning. The one closest to our research is the localized learning modification leading to inserting the local learning step size in the SOM weights update formula, in this way, affecting the SOM's BMUs' neighborhood widths. The local learning step size depends on the stimulus density of the weight vectors (prototypes) of SOM. As it is noticed in [13], a major drawback of the approach is that one has to estimate the generally unknown data distribution corresponding to the mentioned stimulus density, which may lead to numerical instabilities of the control mechanism [13]. Such a drawback does not concern the proposal of the present paper, because in our method, there is no necessity of any data distribution estimation. The second important difference between our technique and the localized learning is that in our method, preliminary data clustering and the resulting inner-cluster variances refer to the input samples in the SOM input space, whereas in case of the localized learning, the local learning step size is determined on the basis of the stimulus density of the weight vectors of SOM, i.e., on the basis of the intrinsic SOM structural information.

In the articles [9,11], an adaptive version of the SOM technique is introduced, in which, the frequency information about the input dataset is employed in the adaptive training process of SOM, i.e., in the adaptive form of the SOM's update formula. Strictly speaking, the frequency of occurrences of data samples in the input data space is utilized as the basis for the SOM's BMUs' neighborhood widths determination. The distinction between the approach from [9,11] and our research is that in the present paper, the information on the density and scattering of the input data samples in the input data space is included in the adaptive training process of SOM, whereas in the study from [9,11], the input data samples' frequencies of occurrences are taken into account, when SOM is being trained, and its lattice is being constructed.

Finally, the paper [1] proposes a Local Adaptive Receptive Fields Self-Organizing Map (LARFSOM). Local models are built by calculating between the output associated with the winning node and the difference vector between the input vector and the weight vector. These models are combined by using a weighted sum to yield the final approximate value. The topology is adapted in a self-organizing way, and the weight vectors are adjusted in a modified unsupervised learning algorithm for supervised problems.

3 Traditional SOM Method

The SOM algorithm provides a non-linear mapping from an original input high-dimensional data space onto a resulting output 2-dimensional map of neurons.

Besides the classical algorithmic description of the SOM method, which is well-known in the existing literature (see, e.g., [3]), an additional mathematical scaffolding has been presented in [4].

According to [4], the results obtained by the SOM method are equivalent to the results delivered by minimizing the following error function with respect to the prototypes w_r and w_s:

$$\begin{aligned} e &= \sum_{r} \sum_{x_i \in V_r} \sum_{s} h_{rs} d^2_{\mathrm{Euc}}\left(x_i,\ w_s\right) \\ &\approx \sum_{r} \sum_{x_i \in V_r} d^2_{\mathrm{Euc}}\left(x_i,\ w_r\right) \\ &\quad + K \sum_{r} \sum_{s \neq r} h_{rs} d^2_{\mathrm{Euc}}\left(w_r,\ w_s\right)\ , \end{aligned} \tag{1}$$

where x_i, $i = 1, \ldots, N$ is the ith input sample in high-dimensional input space, N is the total number of input samples; w_r, $r = 1, \ldots, M$ and w_s, $s = 1, \ldots, M$ are the prototypes of input samples in the grid (the different indeces r and s are used in order to compute the sum of distances between neurons within the SOM grid, including the values of the function h_{rs}); M is the total number of prototypes/neurons in the grid; h_{rs} is a neighborhood function (e.g., the Gaussian kernel) that transforms non-linearly the neuron distances (see [3] for other choices of neighborhood functions); $d_{\mathrm{Euc}}(\cdot,\ \cdot)$ is the Euclidean distance; and V_r is the Voronoi region corresponding to prototype w_r.

The width of the kernel h_{rs} is adapted in each iteration of the algorithm using the rule proposed by [5], i.e.:

$$\sigma\left(t\right) \ = \ \sigma_m \left(\frac{\sigma_f}{\sigma_m}\right)^{\frac{t}{N_{\mathrm{iter}}}}\ , \tag{2}$$

where $\sigma_m \approx \frac{M}{2}$ is typically assumed in the literature (e.g., in [3]), σ_f is the parameter that determines the smoothing degree of the principal curve generated by the SOM algorithm [5], and N_{iter} is the total number of iterations during the SOM training process.

4 A Novel Clustering-Based Adaptive SOM Method

The main proposal of this paper is a method for adaptive SOM training, which is based on the preliminary data analysis, i.e., clustering of the input data samples in the input data space. After the clustering process, one calculates the inner-cluster variances for each of the generated clusters. These values represent the density and scattering of the input data samples, which is, as we claim in our research, the crucial information for the proper adaptation and adjustment of the SOM's lattice to the properties and characteristics of the input dataset. Therefore, the inner-cluster variances are subsequently included in the SOM's exponential update formula (2) from the work [5], and consequently, these values are employed in the SOM's unsupervised training process. In our research, we formulate an assertion that the introduced modification to (2) results in a higher SOM's performance and accuracy, and therefore, it can be recognized as the traditional SOM's enhancement.

The main idea behind the proposed method is the utilization of the information about the data density and scattering in the input data space for improving the training of SOM by making it adaptive and intelligent. This information is included during the setting of the neighborhood widths of SOM's BMUs', in this way, constituting a novel adaptive rule for the training of SOM, which is the main contribution of our paper. The information about the data density and scattering in the input data space is obtained from the measurements of the inner-cluster variance possible to determine after a preliminary data clustering in the input data space.

The entire proposal of the extension to the traditional NeRV method is presented completely and formally in Procedure 1.

Procedure 1. *The clustering-based and density-preserving adaptive SOM method proposed in current paper proceeds as follows:*

Step 1. Perform a clustering of the analyzed dataset in the input high-dimensional space.

Step 2. Compute the inner-cluster variance for each of the clusters according to the following formula:

$$\nu_k = \frac{1}{N_k} \sum_{i=1}^{N_k} d\left(c_k,\, x_i\right) , \tag{3}$$

where ν_k is the inner-cluster variance of the kth cluster, $k = 1, \ldots, K$; K is the number of clusters; N_k is the number of data samples in the kth cluster; $d\left(\cdot,\, \cdot\right)$ is a given suitable dissimilarity measure in the input *high-dimensional space;* c_k *is the centroid of the kth cluster; and* x_i *are the data samples in the input high-dimensional space.*

Step 3. Assign the inner-cluster variances ν_k *to each data sample in the input space:*

$$\nu_i = \nu_k \text{ , such that } x_i \in C_k \text{ ,} \tag{4}$$

where ν_i *is the inner-cluster variance of the ith data sample in the input space,* $i = 1, \ldots, N$*;* N *is the total number of data samples in the input space;* C_k *is the kth cluster in the input space,* $k = 1, \ldots, K$*; and the rest of the notation has been explained previously in this paper.*

Step 4. Include the inner-cluster variances ν_i *in the exponential update formula (2):*

$$\sigma_i\left(\nu_i, t\right) = \left(1 + \nu_i\right) \sigma_m \left(\frac{\sigma_f}{\sigma_m}\right)^{\frac{t}{N_{\text{iter}}}} \text{ ,} \tag{5}$$

where σ_i *is the width of the Gaussian kernel used during the training of SOM for the ith data sample in the input space and the BMU in the SOM grid corresponding to that ith data sample.*

Step 4. Minimize the error function (1) utilizing the novel form of the neighborhood function h_{rs} *including the novel adaptive exponential update formula (5) for* $\sigma_i\left(\nu_i, t\right)$*.*

5 Experiments

In our experimental research, we aimed to verify the effectiveness of the approach introduced in the current paper. All of the experiments have been carried out in two phases, i.e., the input data visualization itself and the *a posteriori* output data clustering, i.e., clustering of the visualized data, or in other words, clustering of the data projected on the SOM grid. The data clustering within the visualization space has been conducted using the weight vectors (prototypes) attached to the neurons in the SOM grid. The results of the *a posteriori* SOM projection clustering have served us as the basis of the comparisons between the introduced technique and the three selected reference data visualization methods.

The experiments have been conducted on real data in the two different research fields: in the field of words visualization and clustering and in the field of speakers visualization and clustering. The first part of the experimental study has been carried out on the large dataset of high-dimensionality (Subsect. 5.3), whereas the second part has been conducted on smaller dataset, but also of high-dimensionality (Subsect. 5.4).

As a result, the scalability of our approach is presented, i.e., the ability to effectively operate on datasets of significantly different size. Note that our method does not increase essentially the computational complexity of the traditional SOM, and it adds only the computational demand of the initial clustering, which is run only once, and it is not repeated during the training of SOM.

Hence, in case of the DBSCAN clustering algorithm, the complexity of the initial clustering is $\mathcal{O}\left(N \ \log N\right)$, does not constraint the scalability property of our technique.

As the reference methods in our empirical study, we have chosen the standard SOM algorithm and the two modified versions of the conventional SOM, i.e., the Time Adaptive SOM (TASOM) and the data visualization approach proposed in [9,11], which will be called throughout this paper as the Frequency-Based SOM (FBSOM).

In case of the speakers' dataset, a graphical illustration of the generated SOMs is provided, whereas in case of the "Bag of Words" dataset, no such illustration is given, because of the high number of data samples in this dataset, which would make such images unclear and unreadable.

5.1 Evaluation Criteria

As the basis of the comparisons between the investigated methods, i.e., as the clustering evaluation criteria, we have used the accuracy rate [6,7] and the uncertainty degree [4,7].

Hence, the following two evaluation criteria have been used:

1. **Accuracy rate.** This evaluation criterion determines the number of correctly assigned samples divided by the total number of samples. Hence, for the entire dataset, the accuracy rate is determined as follows:

$$q \ = \ \frac{N_c}{N} \, , \tag{6}$$

   where N_c is the number of correctly assigned samples, and N is the total number of samples in the entire dataset.
   The accuracy rate q assumes values in the interval $\langle 0, 1 \rangle$, and naturally, greater values are preferred.
2. **Uncertainty degree.** This evaluation criterion determines the number of overlapping samples divided by the total number of samples in a dataset. The samples belonging to the overlapping area are determined on the basis of the ratio of dissimilarities between them and the two nearest clusters centroids. If this ratio is in the interval $\langle 0.9, 1.1 \rangle$, then the corresponding sample is said to be in the overlapping area.
   The uncertainty degree is determined as follows:

$$U_d \ = \ \frac{N_o}{N} \, , \tag{7}$$

   where N_o is the number of overlapping samples in the dataset, and N is the total number of samples in the dataset.
   The uncertainty degree assumes values in the interval $\langle 0, 1 \rangle$, and, smaller values are desired.

5.2 Experimental Setup

In our experimental research, we have utilized the DBSCAN clustering algorithm, because of the important and significant advantages of the algorithm from the point of view of our data analysis, i.e., the automatic clusters' number determination and the capability to handle the non-linearly separable data.

The output data *a posteriori* clustering in the SOM visualization space has been conducted using the standard k-means clustering algorithm.

Each of the investigated methods has been run 50 times, because all of the methods are non-deterministic, and by repeating their executions, we obtain results, which may be recognized as more reliable. The randomness of the methods exists in both phases of our data analysis and processing, i.e., in the data visualization and in the following data clustering.

The values of the accuracy rates and uncertainty degrees in Tables 1 and 2 are computed as the arithmetic averages over all the executed runs of the evaluated methods.

Feature extraction of the textual data investigated in the part of our empirical study demonstrated in Subsect. 5.3 was carried out using the term frequency – inverse document frequency (*tf-idf*) approach.

Features of the speakers' sound signals considered in Subsect. 5.4 have been extracted using a method based on the Discrete Fourier Transform (DFT), which is described in details in [8].

5.3 Words Visualization and Clustering

In the first part of our experimental research, we have utilized excerpts from the "Bag of Words" dataset from the UCI Machine Learning Repository [2].

Our dataset consists of five text collections: Enron E-mail Collection, Neural Information Processing Systems (NIPS) full papers, Daily KOS Blog Entries, New York Times News Articles, PubMed Abstracts. The total number of analyzed words was approximately 10,868,000. On the visualizations generated by the investigated methods, five clusters representing those five text collections in the "Bag of Words" dataset were formed.

Experimental Results. The results of this part of our experiments are reported in Table 1, where the accuracy rates and uncertainty degrees corresponding to each of the evaluated methods are given.

Table 1. Accuracy rates and uncertainty degrees of the words visualization and clustering.

	q	U_d
SOM & k-means	8,175,273/10,868,000 = 0.7522	2,459,195/10,868,000 = 0.2263
TASOM & k-means	8,389,009/10,868,000 = 0.7719	2,304,016/10,868,000 = 0.2120
FBSOM & k-means	9,183,460/10,868,000 = 0.8450	1,523,471/10,868,000 = 0.1402
Proposed SOM & k-means	**9,262,276/10,868,000 = 0.8523**	**1,248,863/10,868,000 = 0.1149**

The results obtained in this part of our experiments have shown a superiority of our approach over the other three data visualization algorithms considered as the benchmark methods. The solution introduced in this paper produced the highest value of the accuracy rate and the lowest value of the uncertainty degree among all the investigated methods.

5.4 Speakers Visualization and Clustering

The speakers visualization and clustering experiment has been conducted on the dataset of sound signals gathered independently by the author of this work.

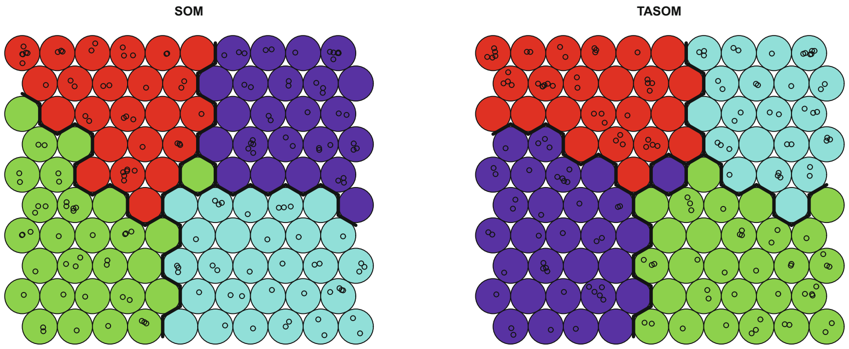
In this part of our experiments, we considered four clusters representing four speakers. Each speaker was represented by 40 speeches, i.e., sound signals (time series). This kind of clustering can be regarded as the speaker recognition based on the sound signals.

Four clusters representing four different speakers have been formed. Each speaker has been represented by 40 10-s sound signals sampled with the 44.1 kHz frequency. Our dataset is composed of 160 sound signals. Feature extraction was carried out according to the DFT-based algorithm, as it was written in Subsect. 5.2. The dataset for the speakers visualization and clustering has been collected autonomously be the author of this research.

Experimental Results. The results of this part of our experiments are reported in Figs. 1a, 1b, 2a, 2b, and in Table 2, which has the same form as Table 1 in Subsect. 5.3. Figures 1a, 1b, 2a, and 2b show the map structures of four considered SOM versions. The points in the 2-dimensional space of these SOM variants' visualizations are the projections of the input data samples from the input data space. In each of the figures, the clusters, generated in the output data *a posteriori* clustering, are indicated and marked with different colors. The colors are assigned to the clusters randomly, therefore, a given cluster may have different colors assigned in different runs of the same algorithm. Hence, the clusters themselves are important, and not their particular colors. Each of Figs. 1a, 1b, 2a, and 2b presents SOM graphics for a single execution of a given SOM version.

Table 2. Accuracy rates and uncertainty degrees of the speakers visualization and clustering.

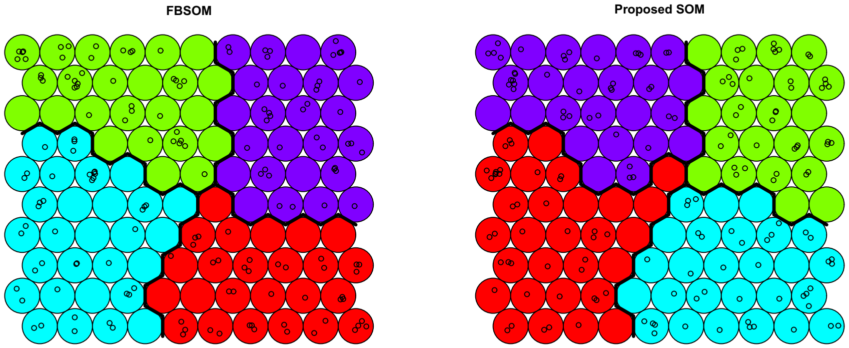
	q	U_d
SOM & k-means	$134/160 = 0.8375$	$15/160 = 0.0938$
TASOM & k-means	$137/160 = 0.8563$	$16/160 = 0.1000$
FBSOM & k-means	$145/160 = 0.9063$	$10/160 = 0.0625$
Proposed SOM & k-means	$\mathbf{154/160 = 0.9625}$	$\mathbf{5/160 = 0.0313}$



(a) Conventional SOM method.

(b) TASOM method.

Fig. 1. Results of speakers visualization and clustering using the conventional SOM method and the TASOM method.



(a) FBSOM method.

(b) Proposed SOM method.

Fig. 2. Results of speakers visualization and clustering using the FBSOM method and the proposed SOM method.

The outcome of the second part of our experiments verified and confirmed the effectiveness and usefulness of our proposed method by indicating that it outperforms all the other tested variants and improvements of SOM. The approach developed in our work returned the higher accuracy rate than all the other algorithms being a subject of evaluation, and furthermore, it allowed for obtaining the lowest value of uncertainty degree, when compared to the reference data visualization methods.

6 Summary

In this paper, a novel version of SOM has been proposed. The novelty in our extension to the traditional SOM was a concept of establishing a relationship between the SOM's BMUs' neighborhood widths and the density and scattering of the data in the input data space. In other words, the SOM's BMUs' neighborhood widths have been determined on the basis of the information about the data density and scattering in the input data space. Precisely speaking, the density and scattering properties of the input data have been numerically represented and conveyed by the quantities of inner-cluster variances obtained as a result of a preliminary input data clustering.

References

1. Ferreira, P.H.M., Araújo, A.F.R.: Growing self-organizing maps for nonlinear time-varying function approximation. Neural Process. Lett. **51**(2), 1689–1714 (2020). https://doi.org/10.1007/s11063-019-10168-9
2. Frank, A., Asuncion, A.: UCI machine learning repository (2010). http://archive.ics.uci.edu/ml
3. Kohonen, T.: Self-Organizing Maps, 3rd edn. Springer, Heidelberg (2001). https://doi.org/10.1007/978-3-642-56927-2
4. Martín-Merino, M., Muñoz, A.: Visualizing asymmetric proximities with SOM and MDS models. Neurocomputing **63**, 171–192 (2005)
5. Mulier, F., Cherkassky, V.: Self-organization as an iterative kernel smoothing process. Neural Comput. **7**(6), 1165–1177 (1995)
6. Olszewski, D.: Asymmetric *k*-means algorithm. In: Dobnikar, A., Lotrič, U., Šter, B. (eds.) ICANNGA 2011. LNCS, vol. 6594, pp. 1–10. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-20267-4_1
7. Olszewski, D.: An experimental study on asymmetric self-organizing Map. In: Yin, H., Wang, W., Rayward-Smith, V. (eds.) IDEAL 2011. LNCS, vol. 6936, pp. 42–49. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23878-9_6
8. Olszewski, D.: *k*-means clustering of asymmetric data. In: Corchado, E., Snášel, V., Abraham, A., Woźniak, M., Graña, M., Cho, S.-B. (eds.) HAIS 2012. LNCS (LNAI), vol. 7208, pp. 243–254. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-28942-2_22
9. Olszewski, D.: An improved adaptive self-organizing map. In: Rutkowski, L., Korytkowski, M., Scherer, R., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (eds.) ICAISC 2014. LNCS (LNAI), vol. 8467, pp. 109–120. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07173-2_11
10. Olszewski, D., Kacprzyk, J., Zadrożny, S.: Time series visualization using asymmetric self-organizing map. In: Tomassini, M., Antonioni, A., Daolio, F., Buesser, P. (eds.) ICANNGA 2013. LNCS, vol. 7824, pp. 40–49. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-37213-1_5
11. Olszewski, D., Kacprzyk, J., Zadrożny, S.: An improved adaptive self-organizing map. In: De Tré, G., Grzegorzewski, P., Kacprzyk, J., Owsiński, J.W., Penczek, W., Zadrożny, S. (eds.) Challenging Problems and Solutions in Intelligent Systems. SCI, vol. 634, pp. 75–102. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-30165-5_5

12. Rauber, A., Merkl, D., Dittenbach, M.: The growing hierarchical self-organizing map: exploratory analysis of high-dimensional data. IEEE Trans. Neural Netw. **13**(6), 1331–1341 (2002)
13. Villmann, T., Claussen, J.C.: Magnification control in self-organizing maps and neural gas. Neural Comput. **18**(2), 446–469 (2006)