An Improved Adaptive Self-Organizing Map

Dominik Olszewski

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

Abstract. We propose a novel adaptive Self-Organizing Map (SOM). In the introduced approach, the SOM neurons' neighborhood widths are computed adaptively using the information about the frequencies of occurrences of input patterns in the input space. The neighborhood widths are determined differently for each neuron in the SOM grid. In this way, the proposed SOM properly visualizes the input data, especially, when there are significant differences in frequencies of occurrences of input patterns. The experimental study on real data, on three different datasets, confirms the effectiveness of the proposed adaptive SOM.

Keywords: Self-Organizing Map, adaptive Self-Organizing Map, neighborhood width, Gaussian kernel, visualization.

1 Introduction

The Self-Organizing Map (SOM) [1] is an example of the artificial neural network architecture. It can be also interpreted as a visualization technique, since the algorithm performs a projection from multidimensional space to 2-dimensional space, this way creating a map structure. The location of points in 2-dimensional grid aims to reflect the similarities between the corresponding objects in multidimensional space. Therefore, the SOM algorithm allows for visualization of relationships between objects in multidimensional space.

The SOM technique is an unsupervised data analysis approach, i.e., there is no additional training data required. Although the method consists of two substantial phases, i.e., the training phase and the testing phase, both of the phases proceed using the same testing dataset. During the training phase, the weights corresponding to each neuron in the SOM grid are being computed. An important step during this process is updating of the neurons in the neighborhood of the Best Matching Unit (BMU) – the closest neuron to the currently matched input pattern. Usually, the neighborhood of the BMU is selected using the Gaussian kernel (see [1] for other choices of neighborhood functions). However, the choice of the neighborhood function parameters, and the choice of the function itself is always to some extent arbitrary, because there are no strict guidelines, and resulting optimal solutions in this matter. Therefore, any justified proposals regarding the neighborhood size of the BMU are desirable, because that choice strongly affects the quality of the final SOM visualization, and consequently, the performance of the entire analysis.

L. Rutkowski et al. (Eds.): ICAISC 2014, Part I, LNAI 8467, pp. 109–120, 2014.
© Springer International Publishing Switzerland 2014

1.1 Our Proposal

In this paper, we propose a method for the SOM neurons' neighborhood widths adaptive computation. The neighborhood widths are determined differently for each neuron in the SOM grid. The introduced method is based on the measurement of the frequencies of occurrences of patterns in the input space. The Gaussian kernel is employed as the neurons' neighborhood function, and the Gaussian standard deviation determining the neurons' neighborhood width is calculated adaptively on the basis of the mentioned frequency. Therefore, the whole considered SOM is an adaptive enhancement to the traditional approach. In case of input patterns appearing frequently in the input space, the corresponding BMU's neighborhood is wider than in case of input patterns occurring rarely in the input space. Consequently, the proposed adaptive SOM reserves larger area for frequent input patterns, and smaller area for rare input patterns. In this way, the novel SOM properly visualizes input data, especially, when there are significant differences in frequencies of occurrences of input patterns in the input space. As a result, the entire visualization comprising the final result will reflect the input data more accurately.

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) [2], the asymmetric SOM [3, 4], the supervised SOM [5], and the adaptive SOM [6–11], to name a few. Naturally, the adaptive SOM versions are of particular interest for the purposes of our research.

In the paper [11], a statistical iterative Gaussian kernel smoothing problem is considered. The authors propose a batch SOM algorithm consisting of two steps. In the first step, the training data are partitioned according to the Voronoi regions of the map unit locations. In the second step, the units are updated by taking weighted centroids of the data falling into the Voronoi regions, with the weighing function given by the neighborhood. The neighborhood width is decreased in each iteration of the algorithm. The difference between the approach from the work [11] and the method developed in our paper is that in [11], the neighborhood width is being constantly decreased exponentially according to the adaptation rule (4) introduced in [11], while in our work, the neighborhood width is adapted to a given dataset depending on the dataset's specific properties.

In the paper [10], an Adaptive Double SOM (ADSOM) is proposed. The constructed map is designed for subsequent clustering analysis without requiring of a priori knowledge about the number of clusters. ADSOM updates its free parameters and allows convergence of its position vectors to a fairly consistent number of clusters provided its initial number of nodes is greater than the expected number of clusters.

The paper [9] proposes a Time Adaptive SOM (TASOM). The work, together with the paper [11], is especially important in the context of our research, because it also introduces a method of neurons neighborhood size adaptive computation.

In the approach proposed in [9], every neuron has its own learning rate and neighborhood size. The difference between the solution from [9] and our method is following. In [9], the adaptation of the neighborhood size results from the "closed-loop" learning of the parameter, i.e., the neighborhood size is updated on the basis of the final quality of visualization (so as to minimize an appropriate error function). Consequently, a learning process is a necessary stage of that analysis. On the other hand, in case of our approach, the neighborhood size is computed in the "open-loop" system, only on the basis of the input dataset analysis (i.e., measurement of frequencies of occurrences of input patterns). No learning process is required, and the method does not rely on the final results of the visualization. Consequently, no additional error function is necessary.

In the work [8], an adaptive hierarchical structure called "Binary Tree TASOM" (BTASOM) is introduced. The considered SOM enhancement resembles a binary natural tree having nodes composed of TASOM networks. The BTASOM is proposed to make TASOM fast and adaptive in the number of its neurons.

The paper [7] proposes an adaptive incremental learning algorithm of the SOM weights. According to the algorithm, the SOM weights are updated incrementally using a higher-order difference equation, which implements a low-pass digital filter.

Finally, in the paper [6], an adaptive GHSOM-based approach (A-GHSOM) is introduced as an effective technique to deal with the anomaly detection problem. As the authors claim, their GHSOM enhancement can adapt on-line to the ever-changing anomaly detection. Consequently, according to the authors, A-GHSOM is superior over the standard GHSOM-based methods, and it provides higher accuracy in identifying intrusions, particularly "unknown" attacks.

3 Traditional Self-Organizing Map

The SOM algorithm provides a non-linear mapping between a high-dimensional original data space and a 2-dimensional map of neurons. The neurons are arranged according to a regular grid, in such a way that the similar vectors in input space are represented by the neurons close in the grid. Therefore, the SOM technique visualizes the data associations in the input high-dimensional space.

It was shown in [12] that the results obtained by the SOM method are equivalent to the results obtained by optimizing the following error function:

$$e(\mathcal{W}) = \sum_{r} \sum_{x_\mu \in V_r} \sum_{s} h_{rs} D(x_\mu,\ w_s) \tag{1}$$

$$\approx \sum_{r} \sum_{x_\mu \in V_r} D(x_\mu,\ w_r) \ + \ K \sum_{r} \sum_{s \neq r} h_{rs} D(w_r,\ w_s)\ , \tag{2}$$

where x_μ are the objects in high-dimensional space, w_r and w_s are the prototypes of objects in the grid, h_{rs} is a neighborhood function (e.g., the Gaussian kernel) that transforms non-linearly the neuron distances (see [1] for other choices of neighborhood functions), $D(\cdot,\ \cdot)$ is the squared Euclidean distance, and V_r is the Voronoi region corresponding to prototype w_r. The number of prototypes is sufficiently large so that $D(x_\mu,\ w_s) \approx D(x_\mu,\ w_r) + D(w_r,\ w_s)$.

According to (2), the SOM error function can be decomposed as the sum of the quantization error and the topological error. The first one minimizes the loss of information, when the input patterns are represented by a set of prototypes. By minimizing the second one, we assure the maximal correlation between the prototype dissimilarities and the corresponding neuron distances, this way assuring the visualization of the data relationships in the input space.

The SOM error function can be optimized by an iterative algorithm consisting of two steps (discussed in [12]). First, a quantization algorithm is executed. This algorithm represents each input pattern by the nearest neighbor prototype. This operation minimizes the first component in (2). Next, the prototypes are arranged along the grid of neurons by minimizing the second component in the error function. This optimization problem can be solved explicitly using the following adaptation rule for each prototype [1]:

$$w_s = \frac{\sum_{r=1}^{M} \sum_{x_\mu \in V_r} h_{rs} x_\mu}{\sum_{r=1}^{M} \sum_{x_\mu \in V_r} h_{rs}}, \tag{3}$$

where M is the number of neurons, and h_{rs} is a neighborhood function (for example, the Gaussian kernel of width $\sigma(t)$). The width of the kernel is adapted in each iteration of the algorithm using the rule proposed by [11], i.e.,

$$\sigma(t) = \sigma_m (\sigma_f / \sigma_m)^{t/N_{iter}}, \tag{4}$$

where $\sigma_m \approx M/2$ is typically assumed in the literature (for example, in [1]), and σ_f is the parameter that determines the smoothing degree of the principal curve generated by the SOM algorithm [11].

4 A Novel Adaptive Self-Organizing Map

In this paper, we propose a novel adaptation rule of the SOM neurons' neighborhood widths. The neighborhood widths are determined differently for each neuron in the SOM grid. The proposed rule employs the exponential update (4) from the work [11], includes the information about the frequencies of occurrences of all input patterns, and consequently, provides a more accurate and effective adaptation process than the rule (4) itself.

The SOM neurons' neighborhood widths are adapted in our research using the Gaussian kernels of the following standard deviation:

$$\sigma_i(|x_i|, t) = \frac{|x_i|}{\max_j(|x_j|)} \sigma_m (\sigma_f / \sigma_m)^{t/N_{iter}}, \tag{5}$$

where x_i, $i = 1, \ldots, n$ is a vector of features representing the ith object in analyzed dataset, $j = 1, \ldots, n$, n is the total number of objects, $|\cdot|$ is the L_1-norm meaning the number of objects given as the argument, and the rest of the notation is explained in (4).

By utilizing the information about the frequencies of occurrences of input patterns, the method proposed in this paper exploits the specific nature and character of a given dataset, and this way, it visualizes the dataset in the SOM grid more accurately by better adjusting to the dataset features and properties.

If the Gaussian kernels specifying the SOM neurons' neighborhood width are fitted to the frequencies of occurrences of input patterns, then the resulting SOM will assign the wider neighborhoods (i.e., the larger area in the SOM grid) to the neurons corresponding to the input patterns appearing more frequently in the input space, and likewise, the obtained SOM will assign the narrower neighborhoods (i.e., the smaller area in the SOM grid) to the neurons corresponding to the input patterns appearing less frequently in the input space.

The desirable consequence of this phenomenon is that the proposed improved adaptive SOM is dataset-dependent, and therefore, it reflects properly the relationships between input patterns, especially if the input dataset is highly diverse with respect to the input patterns' frequencies of occurrences.

5 Experiments

In our experimental study, we have evaluated effectiveness of the proposed improved adaptive SOM technique by conducting the clustering process in the SOM grid obtained using the proposed approach and in the SOM grid returned by a reference method. As the reference method, we have used the traditional time adaptive SOM technique. As the clustering method, we have employed the standard well-known k-means clustering algorithm with the correct number of clusters provided a priori as the input data. Clustering process has been carried out in the 2-dimensional space of the SOM grid. The experimental research aims to ascertain the superiority of the introduced adaptive SOM on the basis of the comparison of the clustering results obtained using the proposed SOM and the classical one. The experiments have been conducted on real data in the three different research fields: in the field of words clustering, in the field of sound signals clustering, and in the field of human heart rhythm signals clustering. The first part of the experimental study has been carried out on the large dataset of high-dimensionality (Subsection 5.3), while the remaining two experimental parts have been conducted on smaller datasets, but also of high-dimensionality (Subsection 5.4 and Subsection 5.5). In this way, one can assess the performance of the investigated methods operating on datasets of different size and nature, and consequently, one can better evaluate the effectiveness of the proposed approach.

The sound signals visualization and clustering was carried out on the piano music recordings, and the human heart rhythm signals analysis was conducted using the ECG recordings derived from the MIT-BIH ECG Databases [13].

In case of the piano music dataset and the ECG recordings dataset, a graphical illustration of the U-matrices generated by SOM is provided, while in case of the "Bag of Words" dataset no such illustration is given, because of the high number of instances in that dataset, which would make such images 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 [4, 14] and the uncertainty degree [4]:

1. **Accuracy rate.** This evaluation criterion determines the number of correctly assigned objects divided by the total number of objects.
   Hence, for the entire dataset, the accuracy rate is determined as follows:
   $$q = \frac{m}{n}, \tag{6}$$
   where m is the number of correctly assigned objects, and n is the total number of objects in the entire dataset.
   The accuracy rates q_i and the accuracy rate q assume values in the interval $\langle 0, 1 \rangle$, and naturally, greater values are preferred.
   The accuracy rate q was used in our experimental study as the main basis of the clustering accuracy comparison of the three investigated approaches.
2. **Uncertainty degree.** This evaluation criterion determines the number of overlapping objects divided by the total number of objects in a dataset. This means, the number of objects, which are in the overlapping area between clusters, divided by the total number of objects. The objects 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 object is said to be in the overlapping area.
   The uncertainty degree is determined as follows:
   $$U_d = \frac{\mu}{n}, \tag{7}$$
   where μ is the number of overlapping objects in the dataset, and n is the total number of objects in the dataset.
   The uncertainty degree assumes values in the interval $\langle 0, 1 \rangle$, and, smaller values are desired.

5.2 Feature Extraction

Features of the time series considered in Subsection 5.4 and Subsection 5.5 have been extracted using a method based on the discrete Fourier transform (DFT), which is described in details in [15].

5.3 Words Visualization and Clustering

In the first part of our experimental study, we have utilized excerpts from the "Bag of Words" dataset from the UCI Machine Learning Repository [16]. It is a high-dimensional dataset of strongly asymmetric nature, especially useful in case of the asymmetric data relationships analysis. It is so, because significant differences in frequencies of occurrences of different words in the entire dataset. Therefore, the experimental investigation on the "Bag of Words" dataset clearly shows the superiority of the proposed asymmetric approach over its traditional symmetric counterpart.

Dataset Description. The "Bag of Words" 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. In the SOM grids generated by the investigated methods, five clusters representing those five text collections in the "Bag of Words" dataset were formed.

Text Feature Extraction. Feature extraction of the textual data investigated in this part of our experimental study was carried out using the term frequency – inverse document frequency (*tf-idf*) approach. The Vector Space Model (VSM) constructed in this way is particularly useful in our research, because it implicitly captures the terms frequency (both: local – document-dependent and global – collection-dependent), which are the source of the hierarchy-based asymmetric relationships in analyzed data (i.e., in this case, between words).

The dimensionality of the analyzed VSM model (i.e., the number of features) was chosen as the minimal length of the vocabularies in the five considered text collections. Consequently, the number of features utilized in this part of our experimental study was 6,906. It was necessary to truncate the longer vocabularies in order to build the data matrix comprising the analyzed VSM model. As a result, not all of the words in the remaining four text collections have been taken into account. Nevertheless, the considered experimental problem remains a high-dimensionality issue, and the number and variety of the words in the analyzed vocabularies makes the problem complex and challenging. Of course, also the highly-asymmetric nature of the investigated dataset is preserved.

Experimental Results. The results of this part of our experiments are reported in Tables 1 and 2, where the accuracy rates corresponding to each investigated approach are presented.

The average (arithmetic average) numbers of words assigned to correct clusters reported in Table 1 and words located in the overlapping areas in Table 2 (in numerators of the ratio fractions) were rounded to the nearest integer values.

The results of this part of our experimental study show that clustering of the SOM grid obtained using the introduced adaptive method outperforms clustering of the SOM grid returned by the standard adaptive approach. The proposed approach leads to the higher clustering accuracy measured on the basis of the

Table 1. Accuracy rates of the words clustering

	q
Traditional adaptive SOM	8,389,009/10,868,000 = 0.7719
Proposed adaptive SOM	9,183,822/10,868,000 = 0.8450

Table 2. Uncertainty degrees of the words clustering

	U_d
Traditional adaptive SOM	2,304,016/10,868,000 = 0.2120
Proposed adaptive SOM	1,523,182/10,868,000 = 0.1402

accuracy rate, and also to the lower clustering uncertainty measured on the basis of the uncertainty degree.

5.4 Piano Music Composer Visualization and Clustering

In this part of our experiments, we considered three clusters representing three piano music composers: Johann Sebastian Bach, Ludwig van Beethoven, and Fryderyk Chopin.

Dataset Description. Each music piece was represented by a 30-seconds sound signal sampled with the 44100 Hz frequency. The entire dataset consisted of 70 sound signals. Feature extraction process was carried out according to the Discrete-Fourier-Transform-based (DFT-based) method described in Subsection 5.2.

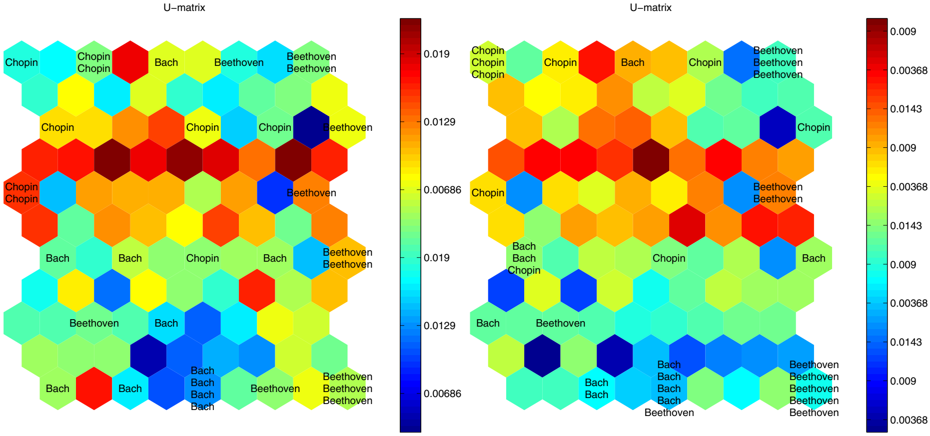
Experimental Results. The results of this part of our experiments are demonstrated in Fig. 1, and in Tables 3 and 4. Figure 1 presents the maps (U-matrices) generated by the symmetric (Fig. 1(a)) and asymmetric (Fig. 1(b)) SOM techniques. The U-matrix is a graphical presentation of SOM. Each entry of the U-matrix corresponds to a neuron in the SOM grid, while value of that entry is the average dissimilarity between the neuron and its neighbors. Table 3, in turn, presents the accuracy rates, while Table 4 reports the uncertainty degrees corresponding to each of the examined approaches.

The average (arithmetic average) numbers of signals assigned to correct clusters reported in Table 3 and signals located in the overlapping areas in Table 4 (in numerators of the ratio fractions) were rounded to the nearest integer values.

Also in this part of our experiments, the proposal of this paper appeared to be superior over the other examined adaptive visualization technique.

5.5 Human Heart Rhythms Visualization and Clustering

The human heart rhythm signals clustering experiment was carried out on the dataset of ECG recordings derived from the MIT-BIH ECG Databases [13].



(a) Traditional adaptive SOM (b) Proposed adaptive SOM

Fig. 1. Piano Music Composers Maps (U-matrices)

Table 3. Accuracy rates of the piano music composer clustering

	q
Traditional adaptive SOM	$27/32 = 0.8438$
Proposed adaptive SOM	$31/32 = 0.9688$

In this part of our experiments, we considered three clusters representing three types of human heart rhythms: normal sinus rhythm, atrial arrhythmia, and ventricular arrhythmia. This kind of clustering can be interpreted as the cardiac arrhythmia detection and recognition based on the ECG recordings.

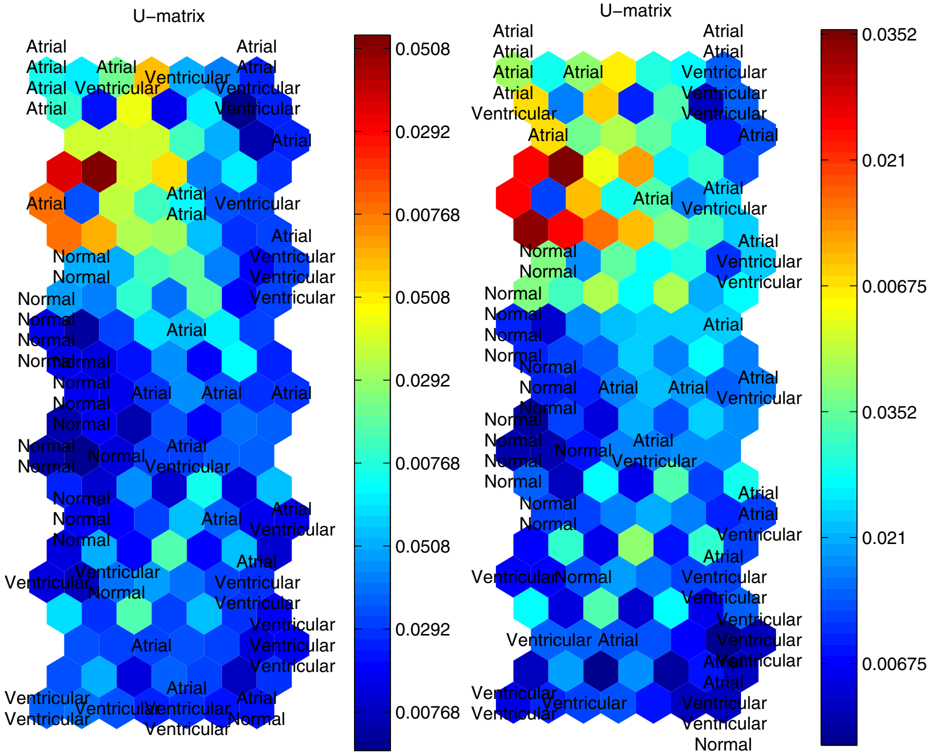
Dataset Description. Our clustering recognizes the normal rhythm, and also, recognizes arrhythmias originating in the atria, and in the ventricles.

We analyzed 20-minutes ECG holter recordings sampled with the 250 Hz frequency. The entire dataset consisted of 63 ECG signals. Feature extraction was carried out according to the DFT-based method described in Subsection 5.2.

Table 4. Uncertainty degrees of the piano music composer clustering

	U_d
Traditional adaptive SOM	$8/32 = 0.2500$
Proposed adaptive SOM	$1/32 = 0.0313$

Experimental Results. The results of this part of our experiments are presented in Fig. 2, and in Tables 5 and 6, which are constructed in the same way as in Subsection 5.4.



(a) Traditional adaptive SOM (b) Proposed adaptive SOM

Fig. 2. Human Heart Rhythms Maps (U-matrices)

Table 5. Accuracy rates of the human heart rhythms clustering

	q
Traditional adaptive SOM	$45/63 = 0.7143$
Proposed adaptive SOM	$58/63 = 0.9206$

Table 6. Uncertainty degrees of the human heart rhythms clustering

	U_d
Traditional adaptive SOM	$18/63 = 0.2857$
Proposed adaptive SOM	$7/63 = 0.1111$

Finally, in the last part of our empirical study, the proposed adaptive SOM clustered by the k-means clustering algorithm produced results superior over the results returned by the reference method clustered using the same algorithm, confirming the usefulness and effectiveness of the proposed solution.

6 Summary

In this paper, a novel adaptive SOM version was proposed. In the introduced approach, the neurons' neighborhood widths are determined using the information about the frequencies of occurrences of input patterns in the input space. The neighborhood widths are determined differently for each neuron in the SOM grid. In case of input patterns appearing frequently in the input space, the neighborhood of the corresponding BMU is wider than in case of the input patterns occurring rarely in the input space. Consequently, the patterns frequent in the input space will receive larger area for their prototypes in the SOM grid, in contrast to the patterns rare in the input space, which will get less place for their prototypes in the grid. In this way, the proposed method provides a proper visualization of the input data, especially, when there are significant differences in the frequencies of occurrences of input patterns, and consequently, our proposal can be regarded as superior over the traditional adaptive SOM technique.

Acknowledgment. This work was prepared as a part of the statutory jobs.

References

1. Kohonen, T.: Self-Organizing Maps, 3rd edn. Springer (2001)
2. Rauber, A., Merkl, D., Dittenbach, M.: The Growing Hierarchical Self-Organizing Map: Exploratory Analysis of High-Dimensional Data. IEEE Transactions on Neural Networks 13(6), 1331–1341 (2002)
3. 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)
4. 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)
5. Płoński, P., Zaremba, K.: Self-Organising Maps for Classification with Metropolis-Hastings Algorithm for Supervision. In: Huang, T., Zeng, Z., Li, C., Leung, C.S. (eds.) ICONIP 2012, Part III. LNCS, vol. 7665, pp. 149–156. Springer, Heidelberg (2012)
6. Ippoliti, D., Zhou, X.: A-GHSOM: An Adaptive Growing Hierarchical Self Organizing Map for Network Anomaly Detection. Journal of Parallel and Distributed Computing 72(12), 1576–1590 (2012)
7. Tucci, M., Raugi, M.: A Filter Based Neuron Model for Adaptive Incremental Learning of Self-Organizing Maps. Neurocomputing 74(11), 1815–1822 (2011)
8. Shah-Hosseini, H.: Binary Tree Time Adaptive Self-Organizing Map. Neurocomputing 74(11), 1823–1839 (2011)
9. Shah-Hosseini, H., Safabakhsh, R.: TASOM: A New Time Adaptive Self-Organizing Map. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 33(2), 271–282 (2003)
10. Ressom, H., Wang, D., Natarajan, P.: Adaptive Double Self-Organizing Maps for Clustering Gene Expression Profiles. Neural Networks 16(5-6), 633–640 (2003)

11. Mulier, F., Cherkassky, V.: Self-Organization as an Iterative Kernel Smoothing Process. Neural Computation 7(6), 1165–1177 (1995)
12. Heskes, T.: Self-Organizing Maps, Vector Quantization, and Mixture Modeling. IEEE Transactions on Neural Networks 12(6), 1299–1305 (2001)
13. Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E.: PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation 101(23), e215–e220 (2000), Circulation Electronic Pages, http://circ.ahajournals.org/cgi/content/full/101/23/e215
14. Olszewski, D.: Asymmetric *k*-Means Algorithm. In: Dobnikar, A., Lotrič, U., Šter, B. (eds.) ICANNGA 2011, Part II. LNCS, vol. 6594, pp. 1–10. Springer, Heidelberg (2011)
15. 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, Part I. LNCS, vol. 7208, pp. 243–254. Springer, Heidelberg (2012)
16. Frank, A., Asuncion, A.: UCI machine learning repository (2010), http://archive.ics.uci.edu/ml