Ubiquitous Self-Organizing Map: Learning Concept-Drifting Data Streams

Bruno Silva[1,⋆] and Nuno Marques[2]

[1] DSI/ESTSetúbal,
Instituto Politécnico de Setúbal, Portugal
bruno.silva@estsetubal.ips.pt
[2] DI/FCT,
Universidade Nova de Lisboa, Portugal
nmm@fct.unl.pt

Abstract. The Internet of Things promises a continuous flow of data where traditional database and data-mining methods cannot be applied. This paper presents a novel variant of the well-known Self-Organized Map (SOM), called Ubiquitous SOM (UbiSOM), that is being tailored for streaming environments. This approach allows ambient intelligence solutions using multidimensional clustering over a continuous data stream to provide continuous exploratory data analysis. The average quantization error over time is used for estimating the learning parameters, allowing the model to retain an indefinite plasticity and to cope with concept drift within a multidimensional stream.

Our experiments show that UbiSOM outperforms other SOM proposals in continuously modeling concept-drifting data streams, converging faster to stable models when the underlying distribution is stationary and reacting accordingly to the nature of the concept-drift in continuous real world data-streams.

Keywords: self-organizing maps, data streams, concept drift, sensor data, clustering, exploratory analysis.

1 Introduction

At present, all kinds of stream data processing based on instantaneous data have become critical issues of Internet, Internet of Things (ubiquitous computing), social networking and other technologies. The massive amounts of data being generated in all these environments push the need for algorithms that can extract knowledge in a readily manner.

Within this increasingly important field of research the application of artificial neural networks to this task remains a fairly unexplored path. The Self-Organizing Map (SOM) [4] is an unsupervised neural-network algorithm with

⋆ This work was partially funded by Fundação para a Ciência e Tecnologia with the PhD scholarship SFRH/BD/49723/2009.

Á. Rocha et al. (eds.), *New Contributions in Information Systems and Technologies*,
Advances in Intelligent Systems and Computing 353, DOI: 10.1007/978-3-319-16486-1_70

topology preservation. The SOM has been applied extensively within fields ranging from engineering sciences to medicine, biology, and economics [5] over the years. The SOM can be visualized as a sheet-like neural network array, whose neurons become specifically tuned to various input vectors (observations) in an orderly fashion. SOM and k-means both represent data in a similar way through prototypes of data, i.e., centroids in k-means and neuron weights in SOM, and their relation and different usages has already been studied [7]. It is the topological ordering of these prototypes in large SOM networks that allows the application of exploratory visualization techniques providing insight on learned data, i.e., clusters and non-linear correlations between features [8].

This paper is the first demonstration of a novel variant of SOM, called *Ubiquitous* SOM (UbiSOM), that is being specially tailored for streaming and big data by using the average quantization error along time to estimate learning parameters. Current SOM variants either estimate parameters based on time which is inadequate for potentially unbounded streams, or the instantaneous output error of the network. This latter approach also suffers from several problems, namely the deficiency to map the input space density, critical for the use of the powerful visualization techniques.

Our experiments show that UbiSOM can be applied to data processing systems that want to use the SOM method to provide a fast response and timely mine valuable information from the data. Indeed our approach, albeit being a single-pass algorithm, outperforms current online and batch SOM proposals in continuously modeling concept-drifting data streams, converging faster to stable models when the underlying distribution is stationary and reacting accordingly to the nature of the concept-drift.

The paper has the following structure: the next section reviews current SOM algorithms that can, in theory, be used for streaming data, highlighting their problems in this setting. The overall methodology of UbiSOM is described in section 3 and experimental results are presented in section 4, by using two artificial datasets and one real world study. The experiments compare the performance over time for stationary and drifting data between UbiSOM and current variants. The real-world application uses sensor data from household electric power consumption. Finally, in section 5 conclusions are drawn and future work is anticipated.

2 Background

A multidimensional stream can be regarded as a continuous, and potentially unbounded, set of observations from a manifold $\Omega \in \mathbb{R}^d$. The SOM establishes a projection from the manifold Ω onto a set of $\mathcal{K}$ neurons, formally written as $\Omega \to \mathcal{K}$. Each neuron i is associated with a prototype $\mathbf{w}_i \in \mathbb{R}^d$, all of which establish the set $\{\mathbf{w}_i\} \in \mathcal{K}$ that is referred as the codebook. The classical *Online* SOM algorithm [4] employs an iterative process between time $t_i = 0$ and time

$t = t_f \in \mathbb{N}^+$ where observations[1] $\mathbf{x} \in \Omega$ are sequentially presented to the map and learning parameters, e.g., learning rate (ε) and neighborhood radius (σ), are decreased monotonically within the time interval $[t_i, t_f]$. Decrease of learning parameters is required for the network to converge steadily towards a topological ordered state and to map the input space density.

However, in a real-world streaming environment t_f is unknown or not defined, so the classical algorithm cannot be used. Even with a bounded stream the Online SOM loses *plasticity* over time and cannot cope easily with changes in the underlying distribution, i.e., *concept drift*. Gama [3] explains why concept drift must be taken into account in a streaming environment: "In data streams the concept about which data is being collected may shift from time to time, each time after some minimum permanence. Changes occur over time. The evidence of drift in a concept is reflected in the observations (e.g., change of mean, variance and/or correlation). Old observations, which reflect the behavior in nature in the past, become irrelevant to the current state of the phenomena under observation".

The proposed SOM algorithm in this paper estimates learning parameters based on the performance of the map over streaming data by monitoring the quantization error (QE) within a sliding window. Some SOM variants have already been proposed to address the parameterization of the SOM not based on time, but based on the local QE, albeit, never intending to process data streams with drifting concepts; the concern was to reduce the parameterization-space and/or accelerate the convergence of the algorithm. The two most recent examples are: the *Parameterless* SOM (PLSOM) [2], which evaluates the local QE and calculates the learning parameters depending on the local quadratic fitting error of the map to the input space, and; the *Dynamic* SOM (DSOM) [6] which follows a similar reasoning by adjusting the magnitude of the learning parameters to the instantaneous QE, but failing to converge from a totally unordered state. Unfortunately, the adjustment of learning parameters is done without evaluating the error in most of neurons. Moreover, authors of both proposals admit that their algorithms are unable to map the input space density onto the prototypes. This has a severe impact on the application of common visualization techniques for exploratory analysis.

3 The Ubiquitous Self-Organizing Map

UbiSOM estimates learning parameters based on the performance of the learning procedure over the most recent observations, trough a sliding window of length T that provides the *average quantization error* $\overline{qe}(t)$ at any particular instant. This provides insight on how well the map is performing on current an recent past data and if drift is occurring. A consequence is that there will always be a learning delay when the underlying concept is truly changing in the order of $T/2$. Nonetheless, it makes the learning more robust to false drifts, so T can be

[1] Normalization of $\mathbf{x}$ is suggested to equate the dynamic ranges along each dimension of $\mathbf{x}$. This ensures that no feature dominates the Euclidean computations, improving the numerical accuracy [4].

seen as a *sensitivity* parameter. This idea stems from a PAC learning model [9] premise which states that "the error rate of the learning algorithm will decrease while the number of examples increase if the distribution is stationary". This idea is empirically described in equation (1), where $\mathbf{w}_c$ is the best matching unit of the map for the observation $\mathbf{x}$ presented at time t.

$$\overline{qe}(t) = \frac{1}{T} \sum_{t}^{t-T+1} \|\mathbf{x}(t) - \mathbf{w}_c(t)\|_{\Omega} \,. \tag{1}$$

Previous works, namely PLSOM and DSOM, use the local quantization error (i.e., the error of the last observation) to estimate learning parameters. However, the local error is very unstable because $\Omega \to \mathcal{K}$ is a many-to-few mapping, where some observations are better represented than others. Since the first value of $\overline{qe}(t)$ is only available at $t = (T-1)$, UbiSOM decreases the learning parameters monotonically as the classical algorithm: $\varepsilon_0 = \varepsilon_i \left(\frac{\varepsilon_f}{\varepsilon_i}\right)^{t/T}$ and $\sigma_0 = \sigma_i \left(\frac{\sigma_f}{\sigma_i}\right)^{t/T}$.

We call this the *bootstrap phase* of the learning procedure which coincides with the ordering phase suggested by Kohonen and T is chosen accordingly to a value never below 1000 [4]. Indeed $T \geq 1000$, has consistently given good results. Then UbiSOM update rule is given by:

$$\Delta\mathbf{w}_i = \varepsilon(\overline{qe}(t))\, h_\sigma(\overline{qe}(t), i, c)\, [\mathbf{x} - \mathbf{w}_i] \,. \tag{2}$$

$$h_\sigma(\overline{qe}(t), i, c) = exp\left(- \left(\frac{\|\mathbf{P}_i - \mathbf{P}_c\|}{\Theta\sigma(\overline{qe}(t))/2} \right)^2 \right), \tag{3}$$

where Θ is a normalization factor related to the lattice size, corresponding to the maximum distance between any two neurons in the lattice. ε and σ are now a function of the average quantization error in time t. In our empirical validation a simple proportion of the average error at time t was then applied to ε_f and σ_f. The learning parameters are increased or decreased proportionally to the variation of the quantization error after $\overline{qe_0} = \overline{qe}(T-1)$, being truncated to ε_f and σ_f, when $\overline{qe}(t) > \overline{qe_0}$.

4 Experimental Results

The presented results show the performance of UbiSOM with stationary and drifting data, using artificial datasets – from which we can establish the ground truth of the expected outcome, and a real-world electric power consumption problem from UCI repository [1], where we further illustrate the potential of UbiSOM when dealing with sensor data in a streaming environment.

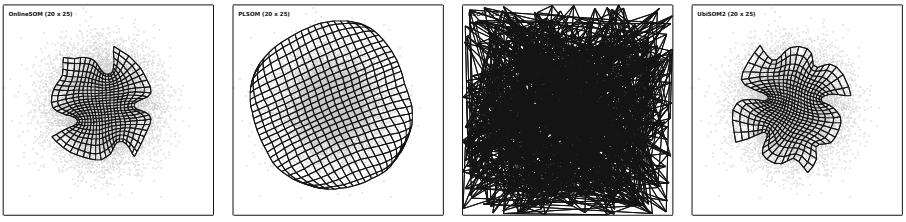
All experiments use a map size of 20×25 with random initial prototypes and input data is normalized, e.g., $\mathbf{x} \in [0,1]^d$. All algorithms are presented with data only once, hence simulating a stream of multidimensional data[2]. After

[2] Given that these data streams are bounded, the classical Online SOM can be correctly parameterized for the presented experiments.

several tries, the best parameters for the chosen map size and for each compared algorithm were selected. The classical *Online* SOM uses $\varepsilon_i = 0.1$ and $sigma_i = \sqrt{k}$, decreasing monotonically to $\varepsilon_f = 0.01$ and $\sigma_f = 1$ respectively; PLSOM uses a single parameter β called *neighborhood range* and the value yielding the best results for the used lattice size was $\beta = 45$. DSOM was initially tried with same parameters as in [6]: $elasticity = 3$ and $\varepsilon = 0.1$, but its convergence from the required completely random initial state (Fig. 1) was not possible and the requirement to initialize the prototypes to evenly cover the input space forced this variant out of further experiments. UbiSOM uses $T = 2000$ (twice the recommended minimum) and parameters: $\varepsilon_i = 0.1$, $\varepsilon_f = 0.08$, $\sigma_i = 0.5$ and $\sigma_f = 0.2$.

4.1 Density Mapping

We illustrate the modeling and quantization of a two-dimensional stream of data (100 000 observations) describing a stationary *Gaussian* distribution, centered in the input space, for all algorithms in Figure 1. It can be seen that only Online SOM and UbiSOM are able to model the input space density correctly, assigning more neurons to the denser area of observations; the later achieves a better convergence. The inability of PLSOM to map the density limits its applicability to exploratory analysis with visualization techniques.

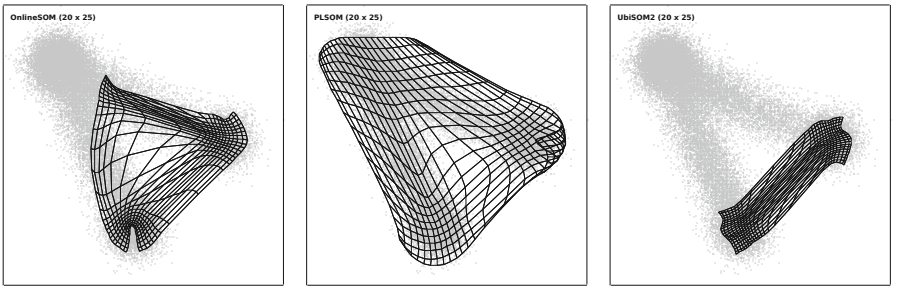


(a) (b) (c) (d)

Fig. 1. Gaussian stream and final maps obtained for: a) Online SOM; b) PLSOM; c) DSOM, and; d) UbiSOM

4.2 Convergence with Stationary and Concept-Drifting Data

The experiments in the presence of a concept that gradually changes over time (drift) are described here. Figure 2 illustrates the trace of dataset drift and displays the final maps obtained for an artificial 2D *Gradual* drift dataset (200 000 observations). The cloud of points starts its drift from the top-left input space and gradually splits into two-clouds in opposite positions. The top cloud moves at linear speed, while the other moves at exponential speed. Only UbiSOM is able to correctly represent the final distribution of the two clouds.



(a) (b) (c)

Fig. 2. Gradual Drift dataset and final maps obtained for (a) Online SOM; (b) PLSOM, and (c) UbiSOM with a single data stream cloud that starts in the top-left quadrant and then splits into two-clouds "drifting" to their final positions

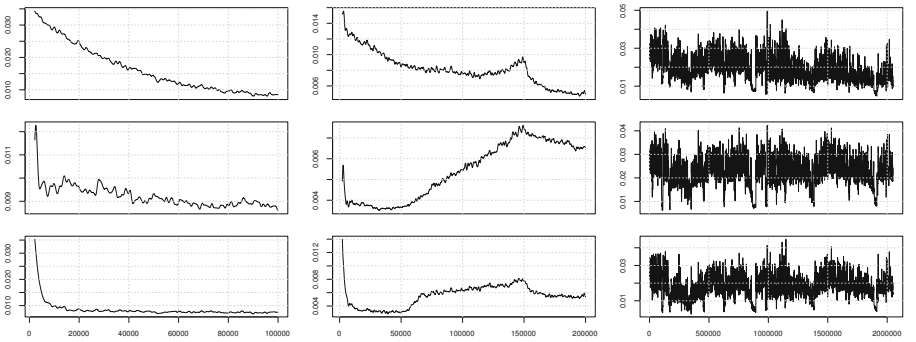


Fig. 3. Average quantization error during learning of different streams for Online SOM (top row), PLSOM (center row) and UbiSOM (bottom row). The columns regard the Gaussian, Gradual and Household datasets respectively.

The average quantization error during learning of different streams was measure for Online SOM, PLSOM and UbiSOM, using the above Gaussian dataset, the Gradual dataset and the real-world *Household* electric power consumption dataset [1], spanning four years (2 049 280 observations with missing data discarded) of collected data to the minute (section 4.3). For easier comparison, all values are dimension-independent, normalized by $(max - min)\sqrt{d}$. The size of the sliding window used to compute the errors is the same that UbiSOM uses, i.e., $T = 2000$. This is considered fair for all algorithms, given that this measure is evaluating their performance over the last T learned observations. Figure 3 depicts the evolution of the average quantization error for all algorithms across the different datasets. In the first dataset (left column) it can clearly be seen that UbiSOM converges faster to a lower average quantization error, which remains stable in this stationary stream. The quantization error of the PLSOM is a little erratic, while the convergence of the Online SOM is dictated by the monotonic

decrease of the learning parameters. Similar observations can be made on the Gradual dataset; moreover, UbiSOM and PLSOM react quickly to the beginning of the gradual drift ($t = 50\,000$), while values of the Online SOM learning parameters at that point in time do not allow it to do so. UbiSOM is able to follow the gradual drift and to converge when the drift stops ($t = 150\,000$); the results are consistent with many other experiments we devised that are not presented due to space constraints. The Household dataset indeed is a continuously-drifting data stream as expected; the overall behavior of the average quantization error seems similar across algorithms, yet, the variance of the errors for the Online SOM is smaller, indicating that it is less reactive to the drifts in the data – this is consistent with the previous dataset.

Table 1 shows the final values at the end of the different streams that corresponds to the final iteration of Figure 3. Results show that comparisons among average quantization error for Online and UbiSOM variants must be evaluated with care, since UbiSOM is accurately describing the final observations presented to the algorithms, while Online SOM tries to describes a static dataset (i.e. in a no-drift scenario with previously known dataset size).

Table 1. Final normalized average quantization errors for the different algorithms in the described datasets. Lower values in bold.

Dataset			Final Average QE		
Name	d	Size	Online SOM	PLSOM	UbiSOM
Gaussian	2	100 000	8.50×10^{-3}	8.61×10^{-3}	$\mathbf{7.48 \times 10^{-3}}$
Gradual Drift	2	200 000	$\mathbf{4.96 \times 10^{-3}}$	6.55×10^{-3}	5.39×10^{-3}
Household	7	2 049 280	$\mathbf{1.07 \times 10^{-2}}$	1.87×10^{-2}	1.47×10^{-2}

4.3 Exploratory Analysis in Real-Time

A real world demonstration is achieved by applying the UbiSOM to the previously mentioned real-world *Household* electric power consumption dataset [1]. Collected data is represented to the minute and only the features relating to sensor values were used resulting in a final dimensionality of the used dataset of $d = 6$. The Household dataset is deemed to contain several drifts in the underlying distribution given the nature of electric power consumption.

Here, we briefly present a visualization technique called *component planes* [8], that further motivates the application of UbiSOM to a concept-drifting data stream. Component planes can be regarded as a "sliced" version of the SOM, showing the distribution of different features values in the map. This visualization can be obtained at any point in time, providing a snapshot of the model for the present and recent past. Ultimately, one can take several snapshots and inspect the evolution of the underlying stream.

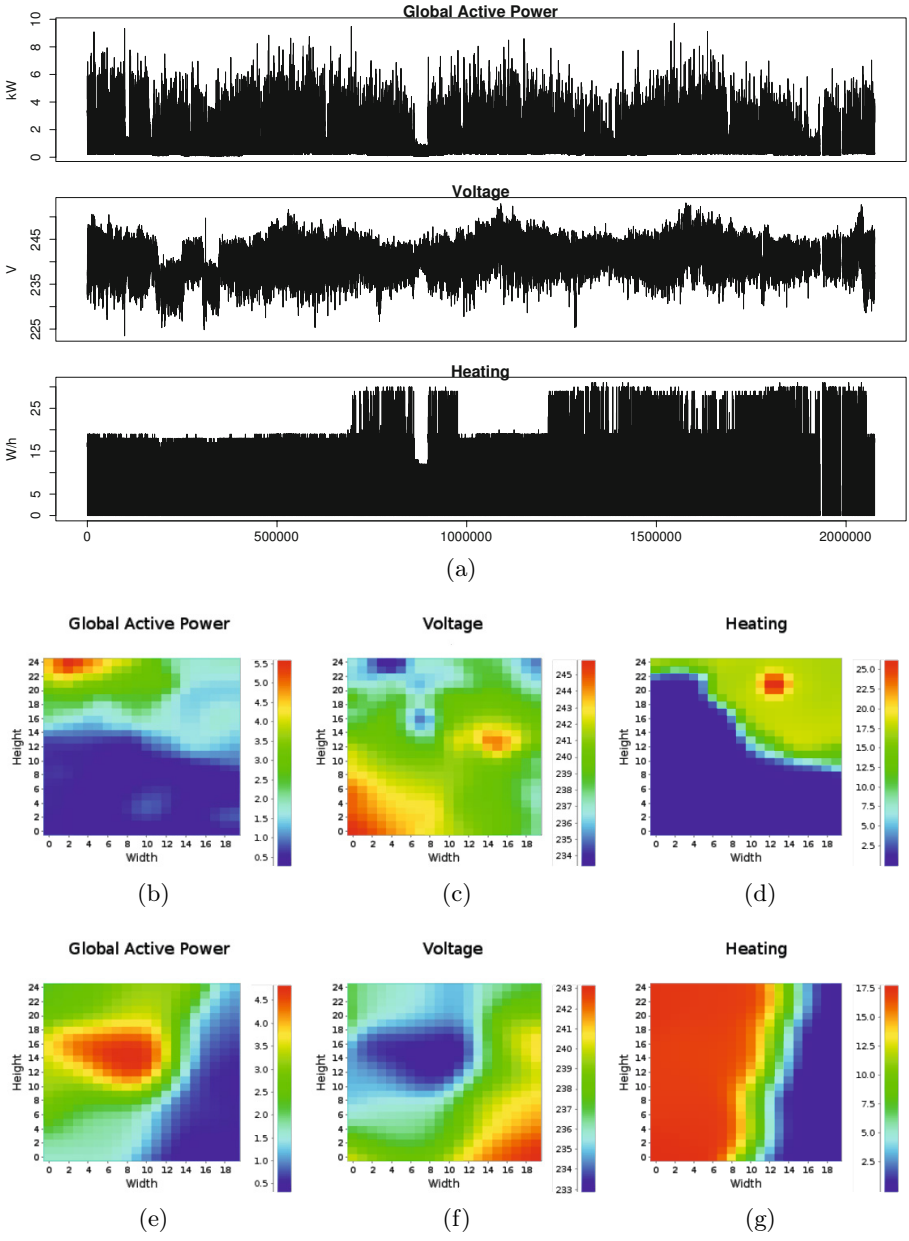


Fig. 4. Household data stream analysis respectively for features: *Global Active Power* (kW); *Voltage* (V) and *Heating* (W/h). (a) shows original values. Illustrative component planes at $t = 1\ 000\ 000$ are shown using Online SOM (b to d) and UbiSOM (e to g).

Figure 4 (e to g) shows some illustrative component planes obtained at the middle ($t = 1\ 000\ 000$) of the entire Household dataset when using UbiSOM. This images indicate correlated features, namely it is visible that the feature *Global Active Power* (kW, in e) is inversely correlated to *Voltage* (V in f). Since UbiSOM is able to map the input space density, the component planes of the heating sensors indicate their relative overall usage in that period of time, e.g., *Heating* (W/h in g) has a high consumption approximately 2/3 of the time. Since this point in time concerns the month of December 2008, this seems self-explanatory.

The comparison of UbiSOM component planes (e to g) with the analogous Online SOM component planes (respectively b to d) is interesting. Indeed, after looking at dataset values for related features in lines (a to c) of Figure 4, observed results confirm the results in Figure 3, with the Online SOM component planes showing less focused and less defined pattern groups that the equivalent component planes of UbiSOM. For example, (a) shows higher values for heating consumption before iteration 1 000 000. While UbiSOM component plane (g) is already representing this, the Online SOM component plane (d) still presents a mix of higher and lower values for this feature. This could be expected since, instead of describing the current data stream, Online SOM tries to describe the full dataset.

5 Conclusions

This paper presented a new SOM algorithm that is being tailored to learn from data streams, called Ubiquitous Self-Organizing Map (UbiSOM). Based on literature review, it is the first SOM variant that is capable of learning stationary and drifting distributions. Experiments presented indicate that the use of the moving average quantization error to estimate learning parameters is a reliable method to achieve the proposed goal and that UbiSOM outperforms current variants in stationary and concept-drifting streams.

UbiSOM usage of average quantization error proportion proved fairly robust. The achieved results show the relevance of the algorithm when applied to data streams. Namely the evolution of the average quantization error showed that UbiSOM is capable of both convergence and reaction to drift. The component-plane based exploratory analysis of the household dataset is particularly relevant for illustrating the behavior of the algorithm over time. Indeed the UbiSOM component planes are more specific and keep the model adapted to distinct usage scenarios. This points to particular useful usage of UbiSOM in many practical applications.

Although theoretical formalization and test of distinct families of functions for quantization error influence are beyond the scope of this paper, presented results motivate its relevance for further ANN studies and models. The value of T should be further explored on distinct settings, since it allows the tunning of model robustness of noise vs. concept drift detection capabilities. Drift detection and signaling can be of great importance, since we can then store current models

that can be later compared to study the evolution of the clusters through the visualization techniques. Applications of high social value where this method may prove useful include using data streams for health monitoring, powering a greener economy in smart cities or financial domains. Ongoing work is now addressing alert systems for strange (or fraudulent) financial data streams based both on macro-economic data and on instantaneous threats to a company resulting from unpredictable market dependencies.

References

1. Bache, K., Lichman, M.: UCI machine learning repository (2013), `http://archive.ics.uci.edu/ml`
2. Berglund, E.: Improved plsom algorithm. Applied Intelligence 32(1), 122–130 (2010)
3. Gama, J., Rodrigues, P.P., Spinosa, E.J., de Carvalho, A.C.P.L.F.: Knowledge discovery from data streams. Chapman & Hall/CRC, Boca Raton (2010)
4. Kohonen, T.: Self-organized formation of topologically correct feature maps. Biological Cybernetics 43(1), 59–69 (1982)
5. Pöllä, M., Honkela, T., Kohonen, T.: Bibliography of self-organizing map (som) papers: 2002-2005 addendum. Neural Computing Surveys (2009)
6. Rougier, N., Boniface, Y.: Dynamic self-organising map. Neurocomputing 74(11), 1840–1847 (2011)
7. Ultsch, A.: Self organizing neural networks perform different from statistical k-means clustering. In: Proceedings of GfKl 1995 (1995)
8. Ultsch, A., Herrmann, L.: The architecture of emergent self-organizing maps to reduce projection errors. In: Verleysen, M. (ed.) Proceedings of the European Symposium on Artificial Neural Networks (ESANN 2005), pp. 1–6 (2005)
9. Valiant, L.G.: A theory of the learnable. Communications of the ACM 27(11), 1134–1142 (1984)