Ubiquitous Self-Organizing Maps

Bruno Silva
DSI/ESTSetúbal
Instituto Politécnico de Setúbal
Portugal
bruno.silva@estsetubal.ips.pt

Nuno Marques
DI/FCT - UNL
Portugal
nmm@di.fct.unl.pt

Knowledge discovery in ubiquitous environments are usually conditioned by the data stream model, e.g., data is potentially infinite, arrives continuously and is subject to concept drift. These factors present additional challenges to standard data mining algorithms. Artificial Neural Networks (ANN) models are still poorly explored in these settings.

State-of-the-art methods to deal with data streams are single-pass modifications of standard algorithms, e.g., *K*-means for clustering, and involve some relaxation of the quality of the results, i.e., since the data cannot be revisited to refine the models, the goal is to achieve good approximations [Gama, 2010]. In [Guha *et al.*, 2003] an improved single pass *k*-means algorithm is proposed. However, *k*-means suffers from the problem that the initial *k* clusters have to be set either randomly or through other methods. This has a strong impact on the quality of the clustering process. CluStream [Aggarwal *et al.*, 2003] is a framework that targets high-dimensional data streams in a two-phased approach, where an online phase produces micro-clusterings of the incoming data, while producing on-demand offline models of data also with *k*-means.

In this position paper we address the use of Self-Organizing Maps (SOM) [Kohonen, 1982] and argue its strengths over current methods and directions to be explored on its adaptation to ubiquitous environments, which involve dynamic estimation of the learning parameters based on measuring concept drift on, usually, non-stationary underlying distributions. In a previous work [Silva and Marques, 2012] we presented a neural network-based framework for data stream mining that explored the two-phased methodology, where the SOM produced offline models. In this paper we advocate the development of a standalone Ubiquitous SOM (UbiSOM), that is capable of producing models in an online fashion, to be integrated in the framework. This allows derived knowledge to be accessible at any time.

The Self-Organizing Map is a well-established data-mining algorithm with hundreds of applications throughout enumerate scientific domains for tasks of classification, clustering and detection of non-linear relationships between features [Oja *et al.*, 2003]. It can be visualized as a sheet-like neural-network array, whose neurons become specifically tuned to various input vectors (observations) in an orderly fashion. The SOM is able to project high-dimensional data onto a 2D lattice, while preserving topological relationships among the input data, thus electing it as a data-mining tool of choice [Vesanto, 1999], either for clustering, data inspection and/or classification. The powerful visualization techniques for SOM models allow the detection of complex cluster structures, detection of non-linear relationships between features and even allow the clustering of time series.

As with most standard data mining methods, classical SOM training algorithms are tailored to revisit the data several times to build good models. As training progresses, existing learning parameters are decreased monotonically over time through one of a variety of decreasing functions. This is required for the network to converge to a topological ordered state and to estimate the input space density. The consequence is that the maps lose plasticity over time, i.e., if a training sample presented at a later point in time is very different from what it has learned so far it does not have the ability to represent this new data appropriately because these parameters do not allow large updates at that time. In ubiquitous environments data is expected to be presented to the network over time, i.e., the network should be learning gradually and derived knowledge should be accessible at any time. This means that the SOM must be able to retain an indefinite *plasticity* over time, with the ability to incorporate very different data from what it has learned at a particular time, i.e., to be in conformance with the "dynamic environment" requirement.

Ubiquitous SOMs , i.e., self-organizing maps tailored for ubiquitous environments with streaming data, should define those parameters not based on time t, but in the error on the network for a particular observation.

An underused variant of the SOM, called the *parameter-less SOM* (PLSOM) [Berglund, 2010]), was first introduced to address the difficulty os estimating the initial learning parameters. The PLSOM has only one parameter β (neighborhood range) that needs to be specified in the beginning of the training process, after which α (learning rate) and σ (neighborhood radius) are estimated dynamically at each iteration. The basic idea behind the PLSOM is that for an input pattern that the network already represents well, there is no need for large adjustments – learning rate and neighborhood radius are kept small. On the other hand, if an input vector is very dissimilar of what was seen previously, then those parameters are adjusted to produce large adjustments. However, in its current form, it fails in mapping the input space density

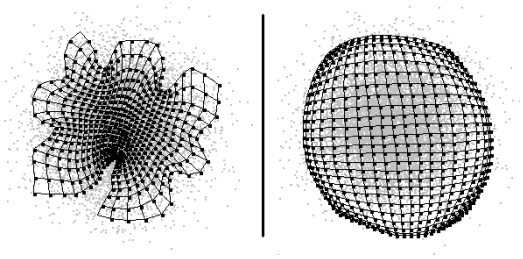

Figure 1: Learned Gaussian distribution for the classical SOM (left) and for the PLSOM (right). The later does not maintain the density of the input space, which undermines the use of visualization techniques for cluster detection and feature correlation.

onto the 2D lattice (Figure 1). This undermines the visualization capabilities of the PLSOM, namely for cluster detection. Also, by estimating the values of the learning parameters solely based on the network error for a given observation, it is very sensible to outliers.

Nevertheless, this variant of the SOM retains an indefinite *plasticity*, which allows the SOM to react to very different input samples from what has been presented to it, at any point in time; and converges faster to an initial global ordered state of the lattice. These two capabilities makes PLSOM an interesting starting point for the proposed goal.

Concept drift means that 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 training samples (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 [Gama, 2010]. In [Silva *et al.*, 2012] we addressed concept drift detection using a different type of neural network, namely Adaptive Resonance Theory (ART) networks. Figure 2 illustrates its applicability to financial time series. It works by measuring the quantization error of the last built micro-cluster ART model, over a predefined number of previous ones. We propose to use the ideas of the PLSOM algorithm using the network error as an input to a concept drift module, either ANN-based or not. While the concept is stable, the learning parameters are being decreased monotonically so as to map the input space density; when the concept begins to drift the parameters are adjusted to higher values so as to cope with the different observations. If the, possibly, underlying non-stationary distribution is drifting rapidly, maintaining higher learning parameters will, consequently, make the model "forget" old and irrelevant observations to the current state.

Conclusion ANN methods exhibit some advantages in ubiquitous data mining: they have the ability to *adapt* to changing environments; have the ability to *generalize* from what they have learned; and through the ANN error it is possible to determine if the new information that arrives is very

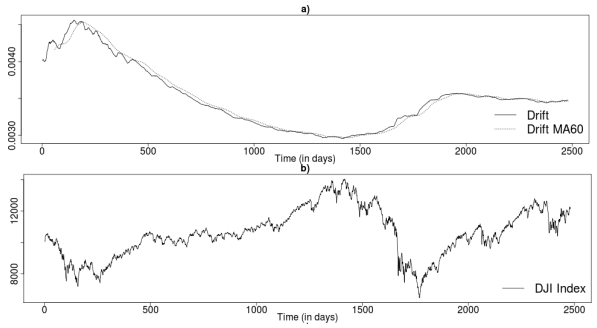

Figure 2: a) Measuring concept drift over a stream of financial statistical indicators in [Silva *et al.*, 2012]. b) The corresponding time series of the Dow Jones Index.

different from what it has learned so far. A purely online Ubiquitous Self-Organizing Map (UbiSOM) that can learn non-stationary distributions is relevant for data stream mining, namely because of its c. The SOM mining capabilities greatly surpass *K*-means, without introducing a big overhead in computation needs. Measuring concept drift as a way to estimate the learning parameters of the learning algorithm is, in our belief, a promising path.

References

[Aggarwal *et al.*, 2003] C.C. Aggarwal, J. Han, J. Wang, and P.S. Yu. A framework for clustering evolving data streams. In *VLDB*, pages 81–92, 2003.

[Berglund, 2010] E. Berglund. Improved PLSOM algorithm. *Applied Intelligence*, 32(1):122–130, 2010.

[Gama, 2010] João Gama. *Knowledge discovery from data streams*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, 2010.

[Guha *et al.*, 2003] S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O'Callaghan. Clustering data streams: Theory and practice. *IEEE Transactions on Knowledge and Data Engineering*, pages 515–528, 2003.

[Kohonen, 1982] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1):59–69, 1982.

[Oja *et al.*, 2003] Merja Oja, Samuel Kaski, and Teuvo Kohonen. Bibliography of self-organizing map (som) papers: 1998-2001, 2003.

[Silva and Marques, 2012] Bruno Silva and Nuno Marques. Neural network-based framework for data stream mining. In *Proceedings of the Sixth Starting AI Researchers' Symposium*. IOS Press, 2012.

[Silva *et al.*, 2012] Bruno Silva, Nuno Marques, and Gisele Panosso. Applying neural networks for concept drift detection in financial markets. In *ECAI2012, Ubiquitous Data Mining Workshop*, 2012.

[Vesanto, 1999] J. Vesanto. SOM-based data visualization methods. *Intelligent-Data-Analysis*, 3:111–26, 1999.