ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning. Bruges (Belgium), 28-30 April 2010, d-side publi., ISBN 2-930307-10-2.

An automated SOM clustering based on data topology

Kadim Taşdemir and Pavel Milenov *

European Commission Joint Research Centre,
Institute for the Protection and Security of the Citizen (IPSC),
Monitoring Agricultural Resources (MARS) Unit
Via E. Fermi 2749, TP266, Ispra, 21027, VA, Italy

Abstract. Self-organizing maps are powerful for cluster extraction due to their ability of obtaining a topologically ordered and adaptive vector quantization of data. Thanks to lower-dimensional representation of high-dimensional data on SOM lattice, clustering is often done interactively from informative SOM visualizations. Yet large volumes of today's data sets necessitate to have automated methods that are as successful as interactive ones for fast and accurate knowledge discovery. An automated SOM clustering, based on hierarchical clustering of a topology representing graph, is proposed here. Applications on several data sets indicate that the proposed method can be successfully used for automated partitioning.

1 Introduction

Self-organizing maps (SOMs) have been widely used for knowledge discovery from high-dimensional data sets such as remote-sensing images and medical imagery, since SOMs enable informative visualizations, which can represent high-dimensional data manifolds on lower-dimensions, to capture detailed cluster extraction. Several innovative visualization schemes have been proposed in the last two decades. For a comprehensive review of these schemes, we refer the reader to [1, 2]. The interactive process required to capture the clusters from SOM visualization, however, needs expert knowledge to interpret the information learned by the SOM and hence is difficult for inexperienced users. This challenge has been considered by introducing automated schemes for cluster extraction without visualization. Commonly used approach is to use hierarchical agglomerative clustering of SOM prototypes, with different distance measures. For example, [3] used Ward's measure, [4] used centroid linkage with SOM lattice neighborhood, whereas [5] used centroid linkage with a gap criterion, and [6] used a recent cluster validity index proposed in [7]. They produce correct partitionings in case of well-separated clusters, but may be unsuccessful for extraction of complex cluster structures. Another approach is to use a recursive flooding of the Clusot surface [8] (a Gaussian surface constructed based on pairwise distances and receptive field sizes of SOM prototypes), however, it produces partitionings similar to that of k-means clustering. A topologically ordered graph clustering [9] is also considered but for visually informative graph representations to describe interactions between objects.

*Figures are in colour, request colour copy by email: kadim.tasdemir@jrc.ec.europa.eu

ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning. Bruges (Belgium), 28-30 April 2010, d-side publi., ISBN 2-930307-10-2.

In this paper, a new scheme for automated SOM clustering is proposed. The method is inspired from the success of interactive clustering based on SOM visualization of a weighted Delaunay graph, CONN [2]. A hierarchical clustering is used due to its advantage of capturing different types of clusters with appropriate measure. Similarities of SOM prototypes are defined by local data distribution within the receptive fields of the prototypes, which is determined by CONN, contrarily to the common usage of distance based similarities. The performance of the proposed method is shown on two synthetic data sets and on a remote sensing image. Section II briefly explains the weighted Delaunay graph CONN and its hierarchical clustering, Section III shows experimental results on three data sets and Section IV concludes the paper.

2 An automated clustering of SOM prototypes

2.1 CONN: A topology representing graph for SOM prototypes

A recent knowledge representation for SOMs is to construct a topology representing graph CONN which is a weighted Delaunay triangulation of the SOM prototypes [2]. CONN indicates how many times two prototypes are selected as best-matching (BMU) and second best-matching unit pair, and is defined as:

$$CONN(i,j) = |RF_{ij}| + |RF_{ji}| \tag{1}$$

where $CONN(i,j)$ is the weight of the edge connecting prototypes p_i and p_j and $|RF_{ij}|$ is the number of data samples in that region of the receptive field of p_i, RF_i, where p_j is the second BMU. CONN, thus, shows the neighborhood relationships of prototypes in the data space with respect to the data manifold: a positive weight between two prototypes ($CONN(i,j) > 0$) indicates similarity (the higher the value the higher the similarity) whereas $CONN(i,j) = 0$ stands for dissimilar prototypes (separated according to the data manifold). Thus CONN represents similarities of prototypes based on detailed local data distribution, contrarily to the common usage of distance based similarities. Interactive clustering from the CONN visualization (rendering CONN on the SOM lattice with lines of various widths and colors) is shown powerful for detailed knowledge discovery [2]. However, as in other SOM visualizations, an interactive process is required to evaluate the visualization (of the line widths and colors) for delineation of cluster boundaries. For fast analysis of large data sets, an automated clustering of the CONN graph is necessary.

2.2 Hierarchical agglomerative clustering of CONN graph

In hierarchical agglomerative clustering of SOM prototypes, each prototype is considered as a singleton cluster and then they are successively merged until a predefined number of clusters is achieved. Several metrics, such as complete linkage, single linkage, average linkage, and Ward's measure, can be used as merging criterion. Among them, complete linkage and single linkage are very sensitive to noise and outliers. Contrarily, average linkage, which merges the two

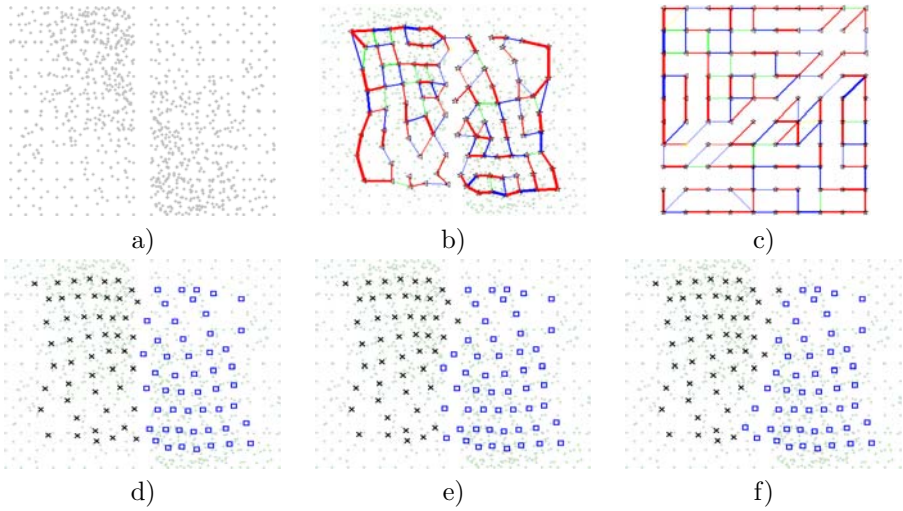

Fig. 1: a) 2-dimensional Wingnut data set [10]. b) CONNvis of SOM prototypes in the data space c) CONNvis on the SOM d) Clustering with CONN based average linkage. Data points are shown by small gray dots whereas the prototypes of two clusters are represented by two different symbols $\mathbf{x}$ and $\square$. Prototypes are correctly partitioned. e) Clustering with distance based average linkage f) k-means clustering.

clusters with the smallest average pairwise dissimilarity, is robust to noise and outliers. Ward's measure or centroid linkage are also insensitive to noise, however, they produce hyper-spherical (or hyper-ellipsoidal) clusters, and in our case of using local density distribution as the dissimilarity measure, they do not have a meaningful interpretation. Therefore, we choose average linkage which considers inter-cluster connectivities according to the data manifold. Since CONN is a similarity measure and hierarchical clustering works with a dissimilarity metric, it is necessary either to convert CONN into a dissimilarity metric by $disCONN(i,j) = max_{i,j}CONN(i,j) - CONN(i,j)$ or revise average linkage to merge the two clusters with the highest average similarity. For this paper, we assume the number of clusters in a dataset is known a priori. Applications in the next section indicate that hierarchical clustering by average linkage based on CONN is successful in the cases where distance based clusterings may fail.

3 Experimental results

We show the performance of the proposed clustering method on 3 data sets. The first one is a 2-dimensional Wingnut data set [10], which has two clusters with inhomogeneous data distribution (Fig. 1.a). CONN visualization (CONNvis) of the 10x10 SOM prototypes is shown in the data space (Fig. 1.b) and on the SOM grid (Fig. 1.c). The boundary between clusters can be visually extracted from these visualizations. Similarly, the proposed automated clustering of the SOM prototypes with average linkage of the CONN graph (Fig. 1.d) finds the two groups correctly, even though the density distribution is inhomogeneous.

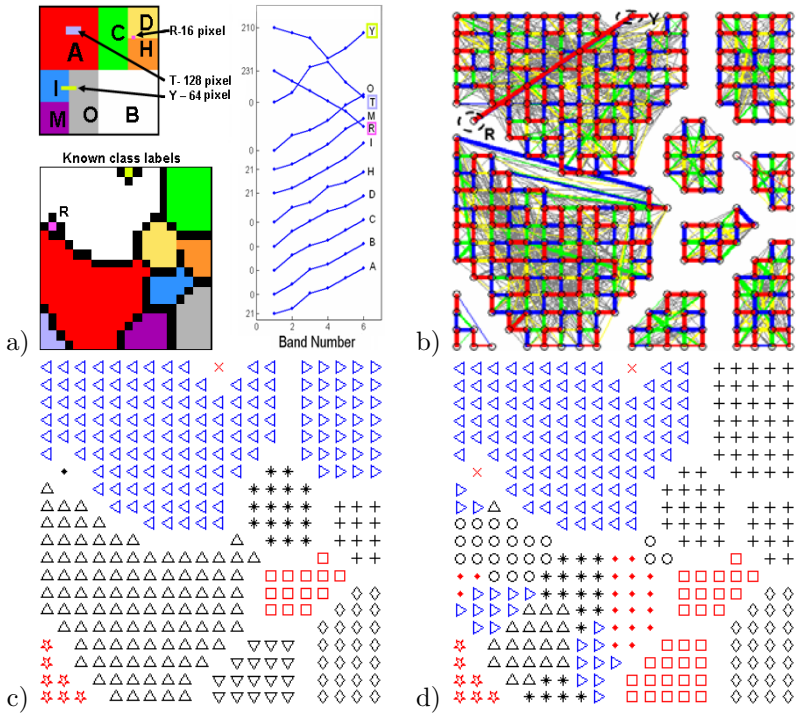

Fig. 2: a) Top left image shows the spatial distribution of the 11 classes of 6-dimensional, 128x128 pixel synthetic spectral image. 3 classes, R, T, and Y are relatively small and different from other 8 classes. Bottom left image indicates the classmap of 20x20 SOM prototypes. Black regions are prototypes with empty receptive fields. The mean signatures of classes are shown on the right, offset for clarity. b) CONNvis of SOM prototypes. All classes can be visually extracted. (Small classes R and Y can be extracted after removing topology violating connection between them [11]) c) Extracted clusters by the proposed method. All 11 classes are extracted correctly. d) k-means clustering. Some clusters are merged into superclusters, whereas cluster A is partitioned into subgroups.

Figs. 1.e and 1.f show two other partitionings, distance based average linkage and k-means clustering, respectively. They cannot find the correct boundary between the two groups, because distance measure, in this case, is not informative enough to capture the clean linear separation between those groups.

The second data set (Fig. 2.a) is a 6-dimensional 128x128 pixel synthetic spectral image with 11 classes (3 of which are rare and significantly different from other 8 classes) [11]. CONNvis of the 20x20 SOM prototypes visually indicate the clear separation between classes. The automated clustering of CONN with average linkage can find all 11 classes correctly. Distance based average linkage is also successful for this data set due to clean separation between clusters, which can be seen from the SOM with existence of prototypes with empty receptive fields between the prototypes of different clusters (Fig. 2.a). However, k-means clustering is unable to capture natural partitions in the data set.

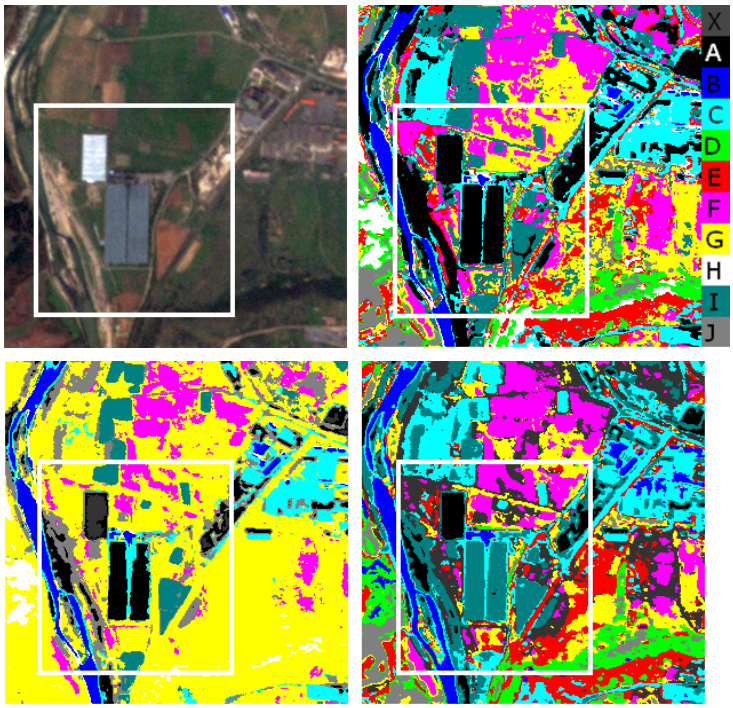

Fig. 3: Cluster maps of a 300x300 subregion of the 2000x2000 remote sensing image. **Top left:** An RGB representation constructed from RapidEye April image with corresponding red, green and blue spectral bands. **Top right:** proposed method: nonagricultural regions (A: urban and constantly bare areas), different kinds of woodlands (D, H and J), water bodies (B), and, various types of agricultural lands (C, E, F, G and I) are correctly clustered. **Bottom left:** distance based average linkage: woodlands and several agricultural clusters are extracted as a supercluster (G, yellow) while there are subclusters (different gray levels) of nonagricultural regions (A, black, on the top right image) **Bottom right:** k-means. Similar to distance based average linkage, clusters are confused. For example, a verified nonagricultural region (grayish twin rectangular areas in the white rectangle on the top left image) is extracted as an agricultural field (I) on the bottom right.

The third data set is a 20-dimensional 2000x2000 remote sensing image of Kardjali, an area in south-eastern Bulgaria. It is a multi-temporal data that is composed of 4 RapidEye images of consecutive months (April to August 2009, each of which has 5 bands: blue, green, red, red edge, near infrared), in order to capture lands that are or can be used for agriculture. A 50x50 SOM is used and 10 different landcover types are obtained by CONN based average linkage. Two other methods (k-means and distance based average linkage) are also used to capture clusters in this image. Due to space constraints, cluster maps for a 300x300 pixel subregion of the 2000x2000 image are shown in Fig. 3. As an initial visual assessment of clustering performance, these cluster maps indicate that the proposed method produces a better partitioning than the other two

methods. For quantitative analysis, 17243 samples are selected. The proposed method has a 96% correct clustering whereas k-means has 93% and distance based average linkage has 50% since some woodland and agriculture clusters are captured as one supercluster (Fig. 3.c). It is also verified by visual interpretation of the four 2000x2000 images and by domain knowledge of a national expert (the second Author) that the proposed method is very successful in determining the boundaries of agricultural regions whereas the other two methods confuse clusters, which in turn produces incorrect land cover identification.

4 Conclusion

A new automated SOM clustering method is proposed in this paper. Even though hierarchical clustering approach has been quite often used in the literature, the definition of prototype (dis)similarities based on data topology (neighborhood relations and detailed local data distribution) is unique. Experiments indicate that it can successfully extract clusters from large remote sensing images and it outperforms the methods that uses distance based similarities. Thus it is promosing for cluster extraction and potentially be a powerful tool in data mining for large data sets. Currently it is assumed that the number of clusters is known a priori. As future work, we plan to decide the number of clusters by using cluster validity indices.

References

[1] J. Vesanto. SOM-based data visualization methods. *Intelligent Data Analysis*, 3(2):111–126, 1999.

[2] K. Taşdemir and E. Merényi. Exploiting data topology in visualization and clustering of Self-Organizing Maps. *IEEE Transactions on Neural Networks*, 20(4):549–562, 2009.

[3] M. Cottrell and P. Rousset. The Kohonen algorithm: A powerful tool for analyzing and representing multidimensional quantitative and qualitative data. In *IWANN 1997 (International Work-Conference on Artificial Neural Networks)*, pages 861–871, 1997.

[4] F. Murtagh. Interpreting the Kohonen self-organizing map using contiguity-constrained clustering. *Pattern Recognition Letters*, 16:399–408, 1995.

[5] J. Vesanto and E. Alhoniemi. Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, 11(3):586–600, May 2000.

[6] S. Wu and W.S. Chow. Clustering of the self-organizing map using a clustering validity index based on inter-cluster and intra-cluster density. *Pattern Recognition*, (37):175–188, 2004.

[7] M. Halkidi and M. Vazirgiannis. A density-based cluster validity approach using multi-representatives. *Pattern Recognition Letters*, (6):773–786, 2008.

[8] D. Brugger, M. Bogdan, and W. Rosenstiel. Automatic cluster detection in Kohonen's SOM. *IEEE Transactions on Neural Networks*, 19(3):442–459, 2008.

[9] Fabrice Rossi and Nathalie Villa. Topologically ordered graph clustering via deterministic annealing. In *Proc. 17th European Symposium on Artificial Neural Networks (ESANN'09), Bruges, Belgium, D-Facto, April 22-24*, 2009.

[10] A. Ultsch. Maps for the visualization of high-dimensional data spaces. In *Proc. 4th Workshop on Self-Organizing Maps (WSOM'03)*, volume 3, pages 225–230, 2003.

[11] K. Taşdemir and E. Merényi. Data topology visualization for the Self-Organizing Maps. In *Proc. 14th European Symposium on Artificial Neural Networks (ESANN'06), Bruges, Belgium, D-Facto, April 26-28*, pages 277–282, 2006.