ESOM Visualizations for Quality Assessment in Clustering

Alfred Ultsch, Martin Behnisch and Jörn Lötsch

Abstract Classical clustering algorithms as well as intrinsic evaluation criteria impose predefined structures onto a data set. If the structures do not fit the data, the clustering will fail and the evaluation criteria will lead to erroneous conclusions. Recently, the abstract U-matrix has been defined for emergent self-organizing maps (ESOM). In this work the abstract forms of the P- and the U* are defined in analogy to the P- and the U*-matrix on ESOM. The abstract U*-matrix can be used for AU*-clustering of data by taking account of density and distance structures. For AU*-clustering the structures seen on the ESOM serve as a supervising quality measure. In this way it can be determined whether an AU*-clustering represents important structures inherent to the high dimensional data. Importantly, AU*-clustering does not impose a geometric cluster shape, which may not fit the underlying data structure, onto the data set. The approach is demonstrated on benchmark data as well as real world data from spatial science.

Keywords Self-organizing maps · U-matrix

A. Ultsch (✉)
DataBionics Research Group, University of Marburg, Hans-Meerwein-Straße, 35032 Marburg, Germany
e-mail: ultsch@Mathematik.Uni-Marburg.de

M. Behnisch
Leibniz Institute of Ecological Urban and Regional Development (IOER), Weberplatz 1, 01217 Dresden, Germany

J. Lötsch
Institute of Clinical Pharmacology, Goethe - University, Theodor-Stern-Kai 7, 60590 Frankfurt Am Main, Germany

J. Lötsch
Fraunhofer Institute of Molecular Biology and Applied Ecology - Project Group Translational Medicine and Pharmacology (IME-TMP), Theodor-Stern-Kai 7, 60590 Frankfurt Am Main, Germany

E. Merényi et al. (eds.), *Advances in Self-Organizing Maps and Learning Vector Quantization*, Advances in Intelligent Systems and Computing 428,
DOI 10.1007/978-3-319-28518-4_3

1 Introduction

It is known that classical clustering algorithms can frequently fail to produce a correct clustering even on data with a clearly defined cluster structure and for which the correct number of clusters is provided as input. This can be demonstrated, for example, on the "Lsun" data set (Fig. 1) from the Fundamental Clustering Problems Suite (FCPS) published as benchmark problems for clustering algorithms [1].

Lsun consists of three clearly separated sets of points on an x-y plane in the form of two elongated rectangular sets forming the letter L and a circular shaped set of points forming the "sun" (Fig. 1, left panel). Popular clustering algorithms such as k-means, Ward, complete- and average linkage all fail to cluster this data set correctly. Figure 1 shows the result of a k-means respectively Ward clustering with the correct number of clusters (i.e. 3) as input (Fig. 1, middle and right panels). The reason for this not uncommon phenomenon of incorrect clustering is that these algorithms imply a geometrical model for the cluster structure. That is, k-means clustering produces a spherical cluster shape, while Ward hierarchical clustering produces a hyperelliptic shape. If this implicit assumption on cluster shape does not fit the underlying data structure, the clustering will fail.

Emergent self-organizing feature maps (ESOM) [2] using the U-matrix [3] represent a topology-preserving mapping of high-dimensional data points $x_i \epsilon R^D$ onto a two-dimensional grid of neurons. In a 3D-display of the U-matrix (e.g. see Fig. 2 in [4]), valleys, ridges and basins indicate a distance-based cluster structure in the data set. Figure 2 (left panel) shows the U-matrix for the Lsun data. The P-matrix on the ESOM enables the visualization of density structures within the data. Both measures, i.e. densities and distances, are combined in the U*-matrix [3] (Figs. 2 and 3). In this way it is possible to discover cluster structures in a data set that are both density- and distance-based. However, ESOM is simply a method to project data from the D-dimensional data space into the plane or the three dimensional landscapes of the

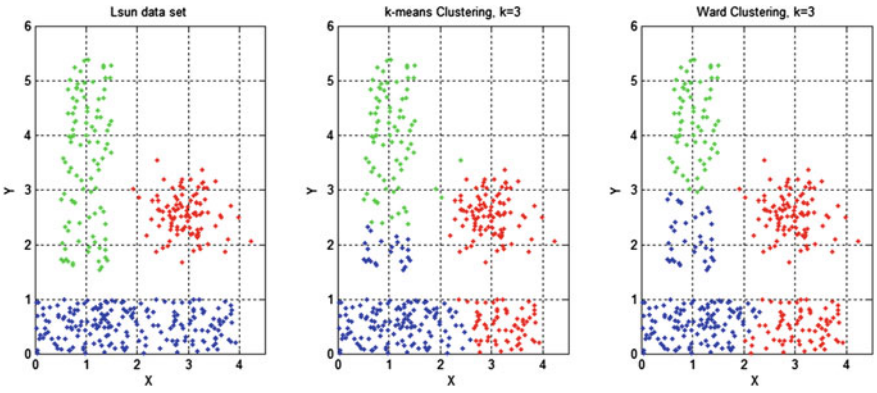

Fig. 1 Lsun data set and some clustering examples

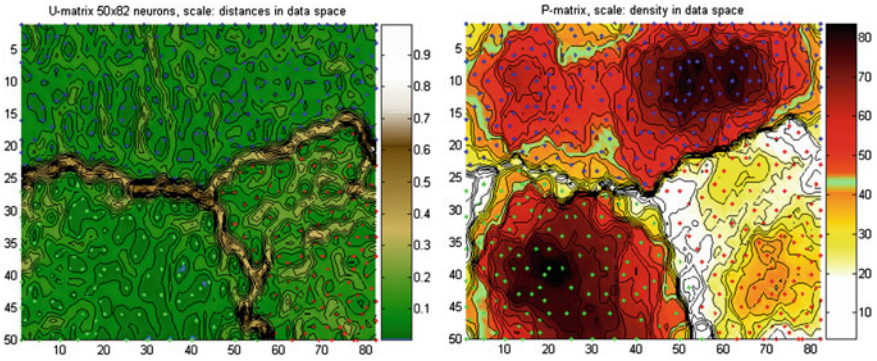

Fig. 2 U- and P-matrix of the Lsun data set

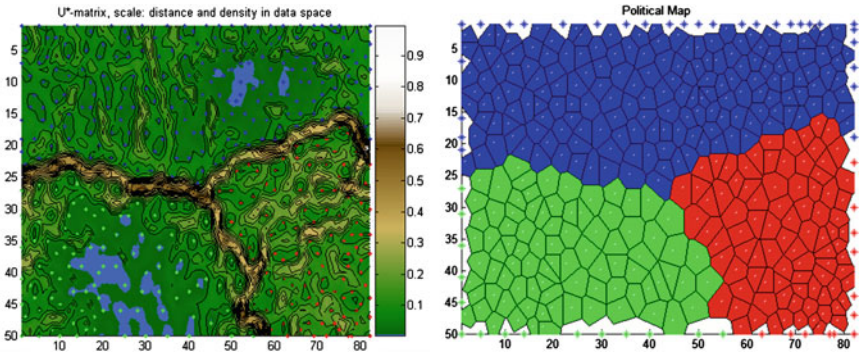

Fig. 3 U*-matrix and "Political Map" for Lsun

U-, P- and U*matrix (ESOM matrices). If cluster structures are revealed through the ESOM matrices, a clustering algorithm is required that can reproduce the structures.

The recently introduced "Abstract U-matrix" (AU-matrix) [5] formally explains the structures seen in the U-matrix. In this work, the abstract P (AP-matrix) and abstract U* (AU*-matrix) are defined. Classical clustering algorithms can be used on the AU*-matrix. The validity of this type of clustering can be assessed by comparing results with the structures seen on ESOM matrices in the form of "Political Maps". The approach is demonstrated on the Lsun data set and on a real-world data set from spatial science research.

2 Methods

The ESOM displays the U-matrix on top of an SOM on the output grid arranged in r rows and c columns using a large ($r^*c > 4000$) number of neurons. Large U-heights in the U-matrix indicate a wide gap in the data space whereas low U-heights indicate

that the points are close to one another. In a 3D display of the U-matrix, valleys, ridges and basins indicate distance-based cluster structures in the data. The P-matrix [3] displays the point density $p(x) = |\{data\ points\ x_i |\ d(x_i, x) <= r\}|$ estimated as the number of data points in a sphere of radius r around x at each grid point on the ESOM's output grid. The U*-matrix combines distance structures (U-matrix) and density structures (P-matrix) into a single matrix (U*-matrix) [3].

The combination can be formalized as pointwise matrix multiplication: $U^* = U^* F(P)$, where $F(P)$ is a matrix of factors $f(p)$ that are determined through a linear function f on the densities $p = p(x)$ of the P-matrix. The function f is calculated such that $f(p) = 1$ if the density p is equal to the median and $f(p) = 0$ if p is equal to the 95-percentile ($p95$) of the densities in the P-matrix. For $p(x) > p95 : f(p) = 0$, which indicates that x is well within a cluster and results in zero heights in the U* matrix. The P-matrix allows the identification of density-based clusters in data sets. The U*-matrix shows a consistent picture of density and distance structures in the data.

The abstract U-matrix (AU-matrix) is a three-dimensional structure with the Voronoi cells of the best-matching units (BMUs) of the data as floor and the data distances corresponding to adjacent Voronoi cells as walls [5]. The AU-matrix can be calculated as the product of the adjacency matrix *Del* of the Delaunay graph of the best-matching units (BMU) with the matrix of distances D between the data points, i.e. $AU = Del * D$. In analogy to the P-matrix, the abstract P-matrix is defined as follows: Let $Del(i, j)$ be an edge in *Del*. This implies that the Voronoi cells of data points x_i and x_j are adjacent. The point (midpoint) $m_{i,j} = mean(x_i, x_j)$ is the point in data space corresponding to $AU(i, j)$. The abstract P-matrix (AP-matrix) contains the densities of all these midpoints: $AP(i, j) = p(mi, j)$. The Abstract U*matrix (AU*-matrix) is calculated in the same way as the U*-matrix (see above). It defines a distance between the data points that takes into account (i) the topology preserving projection of the SOM, (ii) the U-matrix structure and (iii) the density structure of the data. The "Political Map" of an ESOM is a coloring of the Voronoi cells of the BMUs, with different colors for each cluster. Figure 3 (right panel) shows a Political Map for a Ward clustering of the AU*-matrix. A correct clustering using the AU* distances (AU*-clustering) coincides with the structures seen on the ESOM-matrices. Thus, AU*-clustering is a clustering of the data whose results can be visually inspected and supervised using the ESOM-matrices and, in particular, using "Political Maps". This concurs with the structures seen in the other ESOM matrices and enables the validation or invalidation of the data clustering.

3 Relationship to Other Approaches

The Abstract U-matrix (AU-matrix), as well as the extensions presented here (AP-matrix, AU*-matrix), are concepts which help to understand what an empirical U-Matrix, respectively P-Matrix and U*-Matrix, shows which is constructed by the

learning algorithm of an SOM on a data set. The concepts presented here are designed for emergent SOMs (ESOM). These have the property of using SOM which have a very large number of neurons, even substantially more neurons than data points. From our perspective, the number of neurons can be thought of as the pixel resolution of a digital photo camera: the more pixels (neurons) the better the image resolution, i.e. the representation of high dimensional data space. It is clear that time and costs for data processing increase with the number of neurons. However, two factors serve to reduce this burden: improved learning algorithms for the SOMs and Moore's law, which famously states that computing power doubles every two years.

A different approach to Kohonen maps is the so-called k-means-SOM, which uses only few units to represent (clusters of) data. For example, Cottrell and de Bodt use 4×4 units to represent the 150 data points in the Iris data set [6]. In contrast to these approaches, ESOMs represent more of the high dimensional space in their neurons than just the BMUs of the data points. BMUs on ESOM only have more than one data point as attractors if they are practically identical in data space. The connectivity matrix CONN [7–9] assumes non-zero density of data points within the attractor field, i.e. the number of data points projected onto one BMU. The number of data points in these Voronoi cells represents a frequency count. However, this is *not* a valid density measure, since the volumes of the Voronoi cells of different BMUs may be quite different.

A single wall of AU matrix represents the true distance information between two points in data space. A valid density information at the midpoints between BMU and second BMU (notation taken from [7–9]) is calculated for the AP-matrix, since the same volumes, i.e. spheres of a predefined radius, are used. The AU*-matrix therefore represents the true distance information between two points weighted by the true density at the midpoint. The representation is such that high densities shorten the distance and low densities stretch this distance. Using transitive closure for these weighted distances allows classical clustering algorithms (AU*-clustering) to actually perform distance- and density-based clustering, taking into account the complex topology of partially entwined clusters within the data.

As the walls of the AU*-matrix are "paper-thin" there is hardly any way to actually display the AU*-matrix directly. However, an empirical given U*-matrix can and should be adjusted, scaled and normalized to fit best the properties of the AU*-matrix. Such a normalized U*-matrix can then be understood as a visualization of the abstract AU*-matrix.

4 AU*-clustering of the Benchmark Data Set

A top view of the U-matrix using a geographical analogy for color-coding of distances separates the two classes visually as a ridge between valleys (Fig. 2 left panel). This allows the identification of the number of clusters. The P-matrix (Fig. 2 right panel) shows particularly low data densities at those neurons where high values in the

U-matrix are observed. This confirms that the parameter for the density calculation, i.e. the radius of the Parzen window (sphere), is correctly chosen. Furthermore, it shows that the density in the red class (sun) is considerably lower than in the two L-classes in Lsun.

The U*-matrix shown in the left panel of Fig. 3 displays enhanced ridges between the prospective clusters and indicates the cluster centers. The results of the AU*-clustering using Ward clustering on the AU*-matrix are shown as the "Political Map" in Fig. 3. Clustering accuracy using AU*-clustering of the Lsun data was 100 % as compared with the true classification shown in Fig. 1 (left panel).

5 AU*-clustering Applied to FCPS Data Sets

AU*-clustering (AU*C) is the application of a classical clustering algorithm using the AU* distances taken from the Abstract AU*-matrix. Here AU*C-clustering was applied to the data sets in the Fundamental Clustering Problems Suite (FCPS) [10]. FCP was accessed on September 15th, 2015, and downloaded from http://www.uni-marburg.de/fb12/datenbionik/downloads/FCPS.

FCPS offers a variety of clustering problems that any algorithm should be able to handle when facing real world data [10], and thus serves as an elementary benchmark for clustering algorithms. FCPS consists of data sets with known a priori classifications that are to be reproduced by the algorithm. All data sets are intentionally created to be simple, enabling visualization in two or three dimensions. Each data set represents a certain problem that is solved by known clustering algorithms with varying degrees of success. This is done in order to reveal the benefits and shortcomings of the algorithms in question. Standard clustering methods, e.g. single-linkage, ward und k-means, are not able to solve the FCPS problems satisfactorily [10].

Here the accuracy of data clustering, i.e. agreement of U*C on FCPS with the a priori classification, was as follows:

Data Set	Accuracy (%)
Atom	100.00
Chainlink	100.00
EngyTime	95.00
Hepta	100.00
Lsun	100.00
Target	100.00
Tetra	99.00
TwoDiamonds	100.00
WingNut	100.00
GolfBall	100.00

6 AU*-clustering Applied to Spatial Science Data

The AU*-clustering was applied to a data set describing the dynamics of land consumption in all of Germany's municipalities ($n = 11,441$; data valid as of 31.12.2010). The data set captures changes in land consumption in the years 2000 to 2010. Land consumption dynamics (LCD) are described along four dimensions: changes in land usage, changes in population density, changes in trade tax revenues and changes in municipal populations. The rededication of open space into settlement and transportation areas has long been the subject of debate. In many related works, clustering has been employed as a popular method intended to answer specific research questions such as: "How many forms of land consumption exist in Germany?" Most recent approaches have used a Ward or k-means clustering [11, 12]. However, many of these approaches have not validated the clustering. As mentioned above, k-means and Ward clustering algorithms are limited to finding clusters of specific shape, e.g. spherical or ellipsoid respectively for a predefined number of clusters.

The LCD data was ESOM projected onto a grid of $50 \times 160 = 8000$ neurons. Figure 4 shows the U*-matrix of this projection. An AU*-clustering of the data resulted in eight different clusters. Figure 5 shows the political map of this clustering. A comparison with the U* Matrix of the same data set shows excellent coincidence of the observed structures. The ESOM matrices in Figs. 4 and 5 are toroid, i.e. the borders top-bottom and left-right connect to one another [3]. The identified clusters could be related to previously unknown structures of spatial effects in land consumption in German municipalities. For example, one of the clusters indicates that an increase in trade tax per inhabitant was unexpectedly associated with a loss in open spaces and also in population. This points to possible problems in municipal development. Another cluster could be characterized as comprising communities

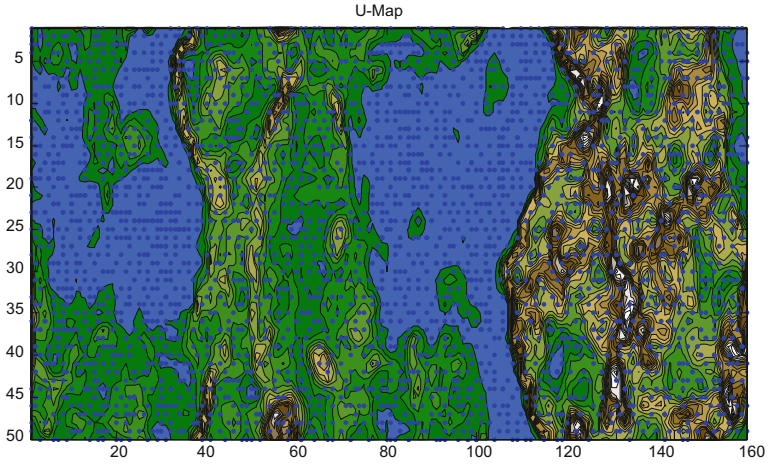

Fig. 4 U*-matrix of the LCD-data set

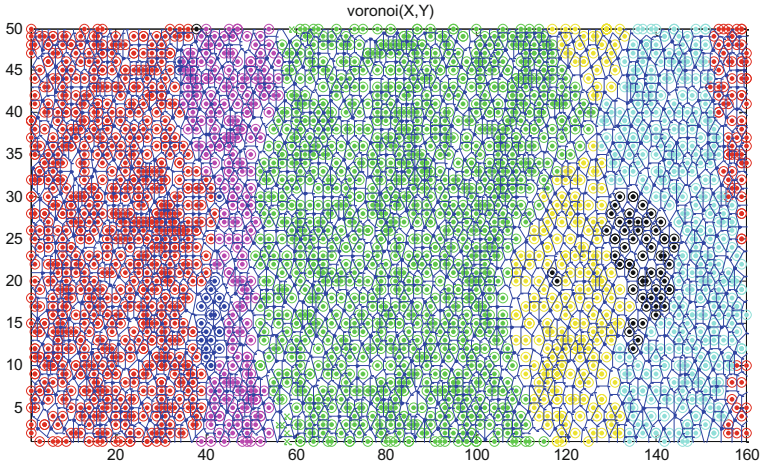

Fig. 5 Political Map of an AU*-clustering of the LCD data set

undergoing the highest change in land consumption within one decade. This could be observed particularly in periurban rural areas. Such results help in the development and optimization of planning programs for sustainable land development. Moreover, the results can be used to help establish a monitoring framework and as the basis for support systems for spatial decision-making. Thus, AU*-clustering offers a deeper multidimensional description of the characteristics of municipal land consumption for cooperating spatial experts.

7 Discussion

Clustering algorithms belong to the class of unsupervised algorithms in Machine Learning. As no desired or "correct" results are available, the results of the algorithm cannot be directly evaluated with respect to their correctness, i.e. no extrinsic evaluation is readily possible. Intrinsic evaluation measures for clustering methods try to capture numeric features of distances with respect to the assumed clusters. They rely on the assignment of low values to the distances within a cluster and of large values to the distances between clusters. However, these measures also implicitly define the geometrical structure of an optimal cluster. For example, the popular silhouette coefficient [13] compares the average distance to elements within the same cluster with the average distance to elements in other clusters. This defines the sphere as the optimal cluster shape. As a consequence, silhouette coefficients do not favor the best cluster structure but rather the cluster structure found by a k-means clustering. Therefore, intrinsic evaluation measures do not allow for the conclusion that some clustering algorithms are better than others as they rely on the existence of the

structure imposed by either algorithm. If the data set in fact contains a differing structure, they will neither provide the correct clustering nor allow the quality of the results to be determined. The Ward and k-means results for the Lsun data set demonstrate this effect (Fig. 1).

ESOM are based on the topology-preserving projection of the data onto the output plane by the underlying SOM. The structures seen on the ESOM matrices therefore allow visual (in-)validation of the cluster structures in the data. Such structures may be defined by distances (U-matrix), densities (P-matrix) or a combination of both (U*-matrix). The abstract form of these three matrices can be used to understand the perceived structures. In this paper, it is proposed that they may be used for clustering (AU*-clustering). The result of a clustering using the AU*-matrix can be compared to the structures seen in the U*-matrix using "Political Maps". This means that if the clustering reproduces the observed structures, it correctly represents (topologically) the structural features of a data set. The algorithm does not impose a model of cluster structure onto the data set. In the data on land consumption dynamics, the AU*-clustering approach produced a map showing eight different types of dynamics. It could be validated with regard to the ESOM matrices constructed for this data set. The resulting clusters were meaningful for the experts in spatial development and planning.

8 Conclusions

Clustering belongs to unsupervised machine learning algorithms for which no "correct" results exist a priori. Classical clustering algorithms and intrinsic evaluation measures of cluster quality impose a predefined structure onto a data set, which can lead to mis-clustering if the imposed structures do not fit the data. By contrast, the here presented professionally constructed ESOM represents a topologically correct projection of the data. The U-Matrix allows visual inspection of distance structures while the P-matrix enables assessment of density structures in the data, and the U*-matrix combines both. In this work the abstract form of these matrices was used for data clustering (AU*-clustering) where the structures seen in the ESOM matrices proofed as a valid quality measure. It can therefore be concluded that this clustering represents important structures in the data without requiring an implicit predefinition of cluster shape or number.

References

1. Ultsch, A.: Clustering with SOM: U*C. Workshop on Self-Organizing Maps, pp. 75–82. Paris (2005)
2. Kohonen, T.: Self-organized formation of topologically correct feature maps. Biol. Cybern. **43**, 59–69 (1982)

3. Ultsch, A.: Maps for Visualization of High-Dimensional Data Spaces. WSOM, pp. 225–230, Kyushu, Japan (2003)
4. Lötsch, J., Ultsch, A.: A machine-learned knowledge discovery method for associating complex phenotypes with complex genotypes. application to pain. J. Biomed. Inf. **46**, 921–928 (2013)
5. Lötsch, J., Ultsch, A.: Exploiting the structures of the U-matrix. In: Villmann, T., Schleif, F.-M., Kaden, M., Lange, M. (eds.) Adv. Intell. Syst. Comput., vol. 295, pp. 248–257. Springer, Heidelberg (2014)
6. Cottrell, M, de Bodt, E.: A Kohonen Map Representation to Avoid Misleading Interpretations, ESANN'96, pp. 103–110. DeFacto, Bruges, Belgium (1996)
7. Tademir, K., Mernyi, E.: Exploiting data topology in visualization and clustering of self-organizing Maps. IEEE Trans. Neural Netw. **20**(4), 549–562 (2009)
8. Merényi, E., Tasdemir, K., Zhang, L.: Learning highly structured manifolds: harnessing the power of SOMs. In: Biehl, M., Hammer, B., Verleysen, M., Villmann, T. (eds.) Similarity Based Clustering. Lecture Notes in Computer Science, LNAI 5400, pp. 138–168. Springer, Berlin (2009)
9. Tasdemir, K., Merényi, E.: A validity index for prototype based clustering of data sets with complex structures. IEEE Trans. Syst. Man Cybern. Part B. 02/2011 **41**(4), 1039–1053 (2011)
10. Ultsch, A.: Clustering with SOM: U*C. In: Proceedings Workshop on Self-Organizing Maps WSOM, pp. 75–82. Paris, France (2005)
11. Kroll, F., Haase, D.A.: Does demographic change affect land use patterns? a case study from Germany. Land Use Policy **27**, 726–737 (2010)
12. Hietel, E., Waldhardt, R., Otte, A.: Analysing land-cover changes in relation to environmental variables in Hesse Germany. Landsc. Ecol. **19**(5), 473–489 (2004). Springer, New York
13. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Comput. Appl. Math. **20**, 53–65 (1987)