

U*-Matrix: a Tool to visualize Clusters in high dimensional Data.

Alfred Ultsch

DataBionics Research Lab, Department of Computer Science
University of Marburg, D-35032 Marburg, Germany
ultsch@informatik.uni-marburg.de

Emergent self organizing feature maps (ESOM) may be regarded as self organized, topology preserving projections of high dimensional data onto a two dimensional map. On top of this ordered floor space an U-Matrix gives insights into the local distance structures of the data set. Using the ESOM/U-Matrix methods for clustering has the advantage of a nonlinear disentanglement of complex cluster structures. Distances inside a cluster are, however, depicted in the same manner as distances between different clusters on an U-Matrix. This may prevent the detection of clusters in some data sets. This is demonstrated on a data set from tumor research. An enhancement of the U-Matrix by taking density information into account is proposed. This leads to a new visualization tool, called U*-Matrix. The U*-Matrix of the tumor data shows structures compatible with a clustering of the data by other algorithms. The combination of distance and density information is expected to be very useful for data mining and knowledge discovery.

1 Introduction

There are two different prototypical SOM usages [1]. The first are SOM where the neurons are identified with clusters in the data space (k-means SOM). The second are SOM where the map space is regarded as a tool for the visualization of the otherwise inaccessible high dimensional data space. A characteristic of these maps is the large number of neurons used. These SOM consist of thousands or tens of thousand neurons. Such SOM allow the emergence of intrinsic structural features of the data space. We call these maps Emergent SOM (ESOM) [2]. ESOM are self organized projections of high dimensional data onto a two dimensional map. The word map is used in a geographical sense, since the SOM algorithm is designed to preserve topological relationships of the input space. The U-Matrix is the canonical tool for the display of the distance structures of the input data on ESOM [3]. U-Matrix methods have also been used for clustering of high dimensional data. See [1] for an overview of the literature.

In some cases there is however, a problem for clustering based on ESOM/U-Matrix: distances inside a cluster are treated in the same manner as distances between data sets of different clusters. In this paper a data set is regarded where a hierarchical cluster algorithm shows the existence of clusters. In the corresponding U-Matrix, however, no cluster structure can be seen. We propose to enhance the U-Matrix visualization by taking density information, in form of the recently introduced P-Matrix, into account [14]. The combination of U-Matrix and P-Matrix leads to a new visualization tool called U*-Matrix. On this U*-Matrix a cluster structure in the data set can be detected. Correspondence analysis shows, that this structure is compatible with a hierarchical clustering of the data.

2 The data set

The data set used in this paper consists of data on malignant brain tumors, called Glioblastoma. The data was provided by J.R. Iglesias-Rozas of Klinikum Stuttgart, Institut für Pathologie. A subset of variables and cases were selected (see [13] for details). The data set used in the following is called "Glio" and consists of 10 variables and 839 cases. The variables are coded on an ordinal scale as positive integers. The coding of the variables is such, that an Euclidian distance between different cases corresponds to the difference in histological findings. No prior tumor classification was given.

3 U-Matrix

ESOM are a self organizing projection from the high dimensional data space onto a grid of neuron locations. The grid of neurons is usually embedded in a two dimensional manifold. This space is called a map with a geographical interpretation in mind. The learning algorithm of the SOM is designed to preserve the neighborhood relationships of the high dimensional space on the map [1]. Therefore the map can be regarded as a roadmap of the data space.

The often used two dimensional plane as map space has the disadvantage that neurons at the borders of the map have very different mapping qualities than neurons in the center of the map. The reason for this is the different number of neighbors of center vs. border neurons. This is important during the learning phase and structures the projection. In many applications important clusters appear in the corners of such a planar map. The embedding into a borderless manifold, such as a torus, avoids such effects [14]. To visualize such toroid maps, four instances of the grid are tiled and displayed adjacently [14]. All figures in the following are such tiled displays.

The U-Matrix is constructed on top of the map. Let n be a neuron on the map, $NN(n)$ be the set of immediate neighbors on the map, $w(n)$ the weight vector associated with neuron n , then

$$U\text{-height}(n) = \sum_{m \in NN(n)} d(w(n) - w(m)), \text{ where } d(x,y) \text{ is the distance used in the SOM algorithm to}$$

construct the map. The U-Matrix is a display of the U-heights on top of the grid positions of the neurons on the map (Utsch Dortmund). An U-Matrix is usually displayed as a grey level picture [15] or as three dimensional landscape [3]. The U-Matrix has become the standard tool for the display of the distance structures of the input data on ESOM [1]. An U-Matrix displays the local distance structure of the data set.

The U-Matrix delivers a “landscape” of the distance relationships of the input data in the data space. Properties of the U-Matrix are:

- the position of the projections of the input data points reflect the topology of the input space, this is inherited from the underlying SOM algorithm
- weight vectors of neurons with large U-heights are very distant from other vectors in the data space
- weight vectors of neurons with small U-heights are surrounded by other vectors in the data space
- projections of the input data points are typically found in depressions
- outliers in the input space are found in „funnels“.
- “mountain ranges“ on a U-Matrix point to cluster boundaries
- „valleys“ on a U-Matrix point to cluster centers

The U-Matrix realizes the emergence of structural features of the distances within the data space. Outliers, as well as possible cluster structures can be recognized for high dimensional data spaces. The proper setting and functioning of the SOM algorithm on the input data can also be visually checked.

Using the ESOM/U-Matrix methods for clustering has the advantage of a nonlinear disentanglement of complex cluster structures. A canonical example for this has been demonstrated by the author on the synthetic “Chainlink” data set [16]. This data set consists of two clusters of data situated inside two intertwined rings. Figure 1 shows the data and corresponding U-Matrix.

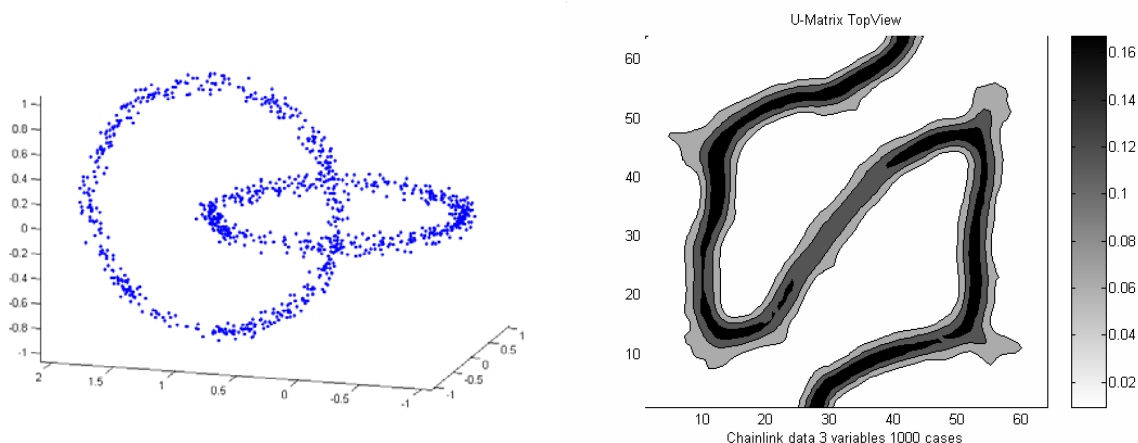


Figure 1: The Chainlink data set and its U-Matrix

The U-Matrix of the Chainlink data shows the disentangling of the data set into two different clusters. Figure 1 right side shows this. Clustering with ESOM/U-Matrix have therefore clear advantages over other clustering algorithms [16].

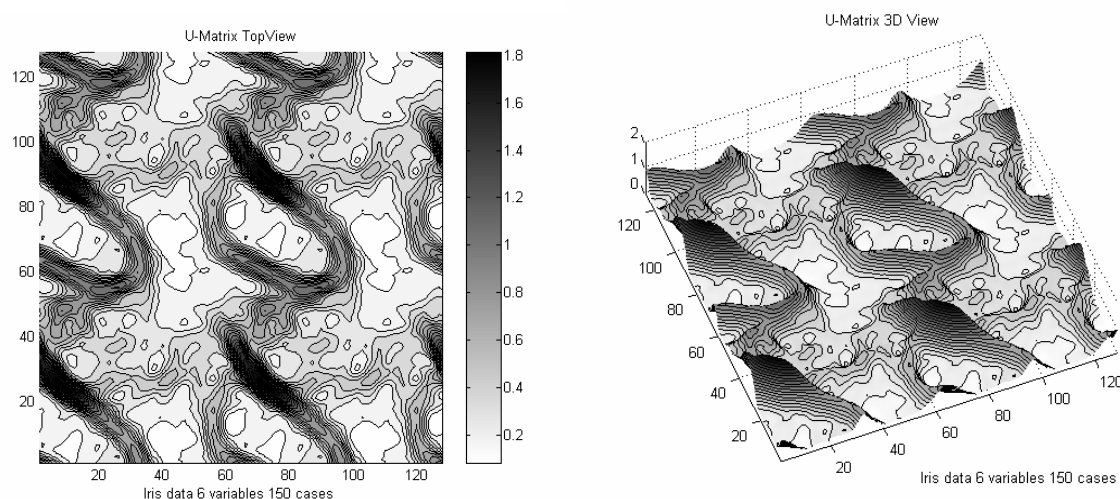


Figure 2: Tiled display of an U-Matrix for the Iris data on 64 by 64 ESOM

In Figure 2 an U-Matrix of the well known Iris data set [17] is shown. In the U-Matrix of the Iris data clusters can be seen as valleys, cluster borders as ridges. The Rand index can be used for comparing two classifications[18]. This index gives a probability of agreement of 86% between the original and the U-Matrix clustering. A comparison of the original classification to a WARD clustering gives a Rand index of 85%. This means that an U-Matrix clustering of the Iris data coincides with the prior classification at least as much as a clustering with the WARD hierarchical clustering algorithm.

U-Matrices have been used in a number of applications to detect new and meaningful knowledge in data sets. To name a few: sea level prediction [5], DNA microarray analysis [4], customer segmentation in mobile phone markets [8], stock portfolio selection [9], and many more[1].

4 A difficult clustering problem for U-Matrix methods

The U-Matrix shows the local distance structure of a topology preserving projection of a high dimensional data set. This can be sufficient to detect cluster boundaries. The Iris data shown above is a typical example. There are, however, cases where a display of local distance structures is not enough to detect clusters.

An agglomerative hierarchical clustering algorithm with the WARD cluster distance function (WARD) was applied to the Glio data. The primary output of a hierarchical clustering algorithm is a dendrogram ordering the cluster distances in form of a tree. Figure 3 shows the dendrogram produced by WARD.

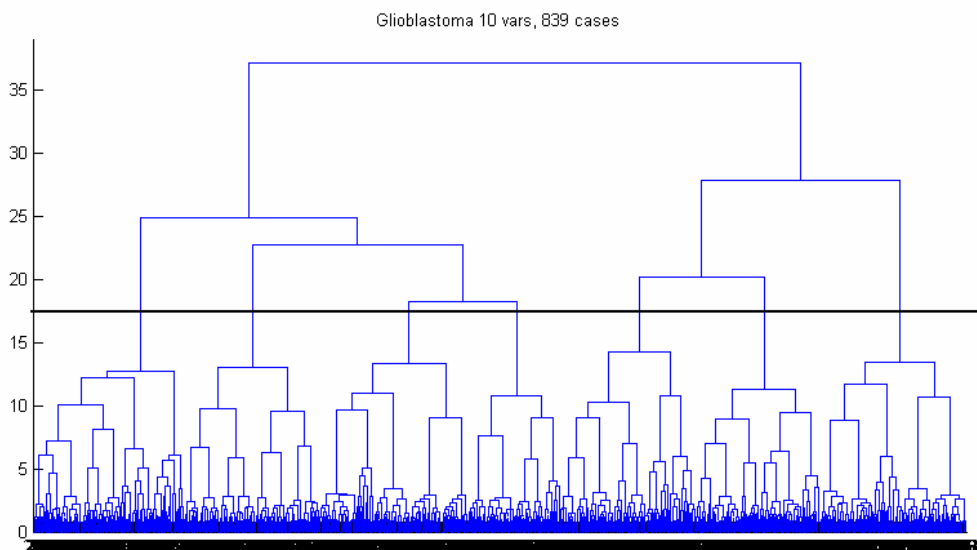


Figure 3: WARD clustering of Glio

In this dendrogram a cluster structure can be seen. Allowing the clusters with a maximum diameter of about 15, leads to the horizontal cutting line shown in Figure 3. This produces seven different clusters for the WARD algorithm on the Glio data set. This cluster structure is, however not at all visible in an U-Matrix of Glio.

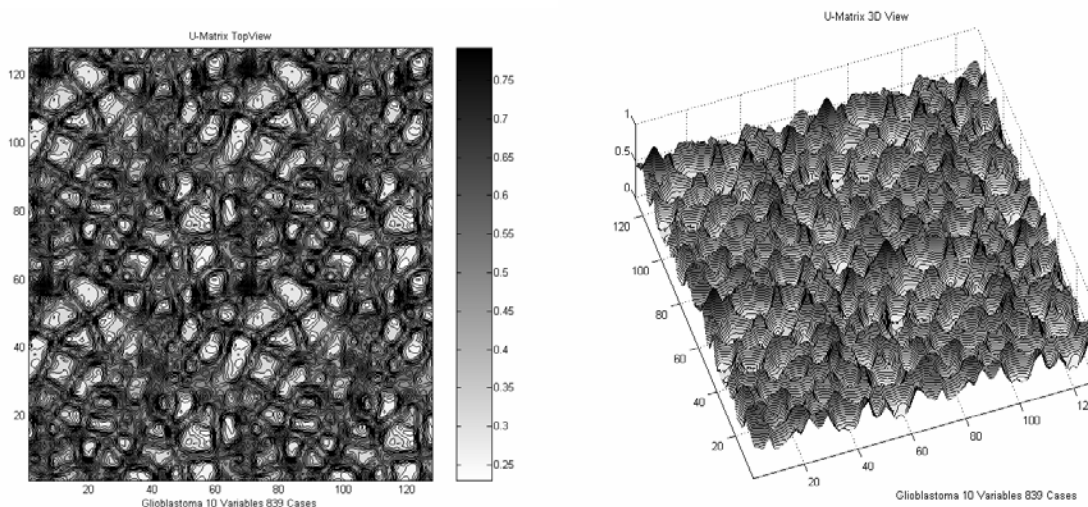


Figure 4 U-Matrix of the Glio data set

Figure 4 depicts a tiled U-Matrix on a 64 by 64 toroid neuronal grid of the Glio data set. There are many “valleys” containing between 10 and 30 projections of data points. The “hills” are about the same height. From this U-Matrix, the existence of clusters can hardly be inferred.

In the next chapters it is shown, how ESOM/U-Matrix methods can be enhanced to detect meaningful clusters in such data sets.

5 P-Matrix

Recently the P-Matrix has been introduced [7]. The P-Matrix also uses the ESOM map as floor space layout. This makes P-Matrix compatible with an U-Matrix. Instead of the local distances, however, density values in data space measured at the neuron’s weights are used as height values.

The P-height of a neuron n , with associated weight vector $w(n)$, is defined as:

$$\text{P-height}(n) = p(w(n), X),$$

where $p(x, X)$ is an empirical density estimation at point x in the data space X .

For each neuron n of a ESOM the P-Matrix displays the density measured in the data space at point $w(n)$, where $w(n)$ is the weight vector associated with neuron n of the ESOM. In principle any density estimation, which works for the input data set of the ESOM can be used. See [10] for an overview on density estimations for multivariate data. Here we use a density estimation called Pareto Density Estimation (PDE) [19]. PDE calculates the density estimation at some point x as the number of points inside a hypersphere (Pareto sphere) around x . The radius of the hyperspheres is called the Pareto radius. The Pareto radius is derived from information optimal sets see [11] for details. It has been shown that PDE leads to a meaningful density estimation and fits nicely into the ESOM /U-Matrix calculation see [12].

A P-Matrix is defined in the same manner as an U-Matrix. The U-Matrix reveals the (local) distance structures, while the P-Matrix gives insights into the density structures of a high dimensional data set. The elements of a P-Matrix are called P-heights. These are the number of points inside the Pareto sphere

In the P-Matrix the density differences in the Glio data set can be clearly seen. Note that the structures are repeated four times due to the tiled mode of display [14].

Properties of a P-Matrix are:

- the position of the projections of the data on the ESOM reflect the topology of the input space, this is inherited from the underlying SOM algorithm
- neurons with large P-heights are situated in dense regions of the data space
- neurons with small P-height are “lonesome” in the data space
- outliers in the input space are found in „funnels“.
- “ditches” on a P-Matrix point to cluster boundaries
- „plateaus“ on a P-Matrix point to regions with equal densities

One can see, that many, but not all, properties of the P-Matrix are the inverse of an U-Matrix display. In contrast to the U-Matrix, which is based on the distance structure of the data space, the P-Matrix is based on the data’s density structure. This gives a new and complementary insight into the high dimensional data space.

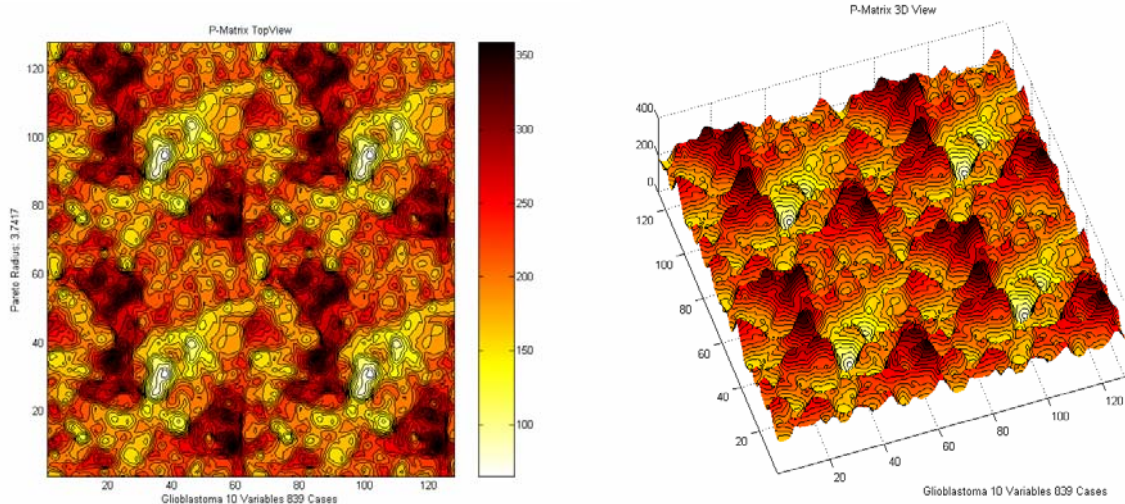


Figure 4: P-Matrix of the Glio data set

Figure 5 shows The P-Matrix of the Glio data set. It can be seen that there are regions where the data is very dense (the dark regions). In these regions substantially more than 200 points can be counted in the Pareto spheres. In other regions are less than 80 points in the Pareto spheres.

6 U*-Matrix

In dense regions of the data space the local distances depicted in an U-Matrix are presumably distances measured inside a cluster. Such distances may be disregarded for the purpose of clustering. In thin populated regions of the data space, however, the distances matter. In this case the U-Matrix heights correspond to cluster boundaries. This leads to the definition of an U*-Matrix which combines the distance based U-Matrix and the density based P-Matrix.

The U*-Matrix is derived from an U-Matrix following these lines:

- when the data density around a weight vector of a neuron is equal to the average data density, the heights shown in an U*-Matrix should be the same as in the corresponding U-Matrix.
- when the data density around a weight vector of a neuron is big, local distances are primarily distances inside a cluster. In this case the U*-Matrix heights should be low.
- when the data density around a weight vector of a neuron is lower than average, local distances are primarily distances at a border of a cluster. In this case the U*-Matrix heights should be higher than the corresponding U-height.

This leads to the following formula: let U-height(n) denote the U-height of a neuron n, mean(P) denote the mean of all P-heights, min(P) the minimum of all P-heights, then the U*-height of an U-Matrix for neuron n, U*-height(n), is calculated as:

$U^*\text{-height}(n) = U\text{-height}(n) * \text{ScaleFactor}(n)$, with

$$\text{ScaleFactor}(n) = \frac{P\text{-height}(n) - \text{mean}(P)}{\text{mean}(P) - \text{min}(P)} \quad (i)$$

From this definition the following is evident:

- $P\text{-height}(n) = \text{mean}(P\text{-heights}) \Rightarrow U^*\text{-height}(n) = U\text{-height}(n)$
- $P\text{-height}(n) < \text{mean}(P\text{-heights}) \Rightarrow U^*\text{-height}(n) > U\text{-height}(n)$ (inter cluster)
- $P\text{-height}(n) > \text{mean}(P\text{-heights}) \Rightarrow U^*\text{-height}(n) > U\text{-height}(n)$ (intra cluster)
- $P\text{-height}(n) = \text{min}(P\text{-heights}) \Rightarrow U^*\text{-height}(n) = 0$ (intra cluster)

A U*-Matrix of Glio can be seen in the following figure:

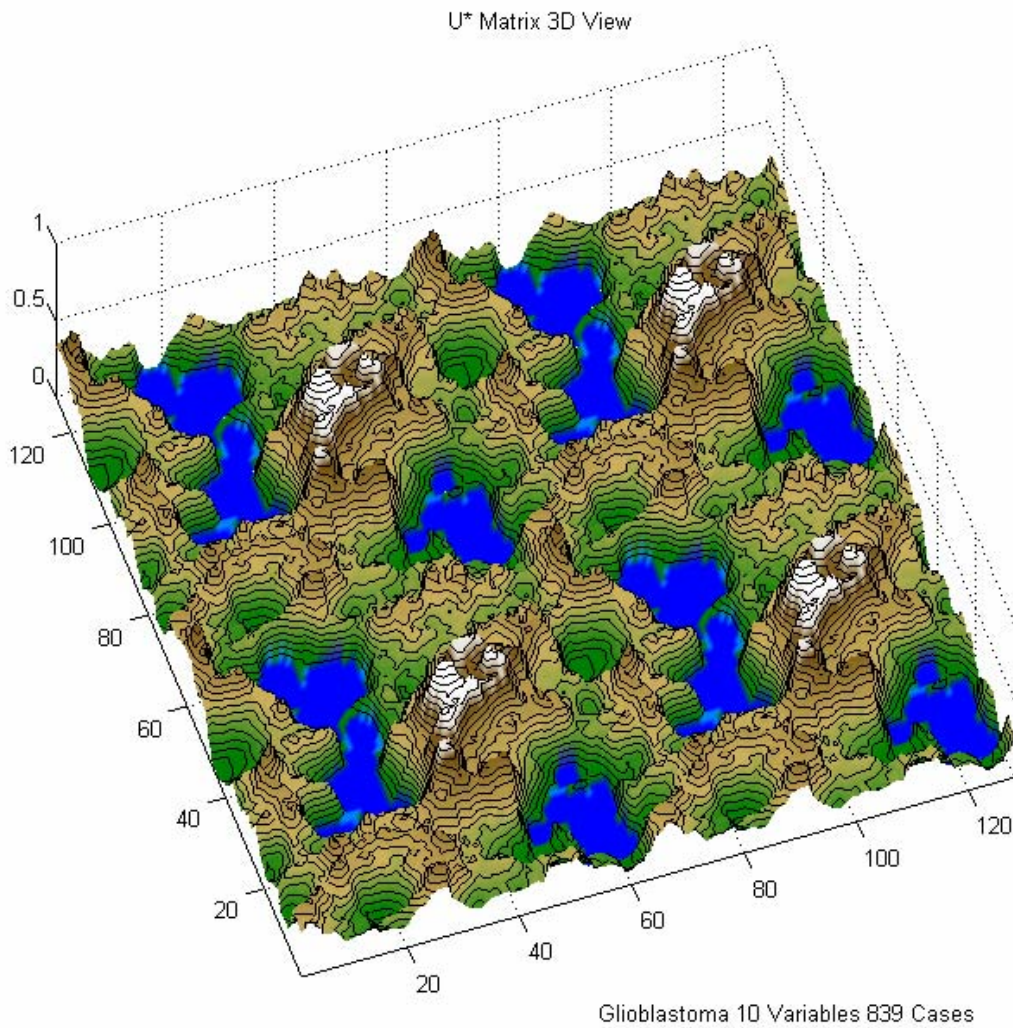


Figure 6: U*-Matrix of the Glio data

The U*-Matrix of Glio exhibits more structure of the 10 dimensional data set than the U-Matrix. In the U*-Matrix clusters in the data set can be identified. In this case a close inspection of the U*-Matrix' structures leads to the identification of 8 classes. The following table allows a comparison of the clustering with the WARD algorithm (WARD classes) and a clustering using the U*-Matrix (U* classes).

U*-Matrix Class\ WARD Class	1	2	3	4	5	6	7	8
1	0%	0%	0%	79%	0%	0%	0%	0%
2	0%	80%	0%	0%	18%	0%	0%	0%
3	4%	0%	40%	0%	23%	1%	17%	1%
4	0%	0%	0%	82%	0%	0%	5%	0%
5	0%	0%	0%	0%	100%	0%	0%	0%
6	44%	1%	1%	0%	2%	49%	1%	0%
7	31%	0%	0%	0%	0%	0%	0%	59%

Table 1: Percentages of WARD classification contained in U*-Matrix classification

One can see that, for example, about 80% of WARD class number 1 is contained in U*-Matrix class number 4. The remaining 20% are not classified by the U*-Matrix classification. For WARD classes 1,2 and 4 there is a corresponding U*-Matrix class. The class number 5 is identical in both cluster algorithms. The remaining WARD classes are split to two or three U*-Matrix classes. The Rand index, giving the probability of agreement of the two clustering algorithms, results to 89%. Compared to values of the Iris data above this indicates a strong coincidence of the two clustering algorithms.

In summary, this experiment shows that a clustering using an U*-Matrix does not contradict a WARD clustering. Similar results can be obtained using other clustering algorithms. Differences in the classifications may be explained by the nonlinear disentangling properties of the ESOM algorithm.

7 Discussion

The detection of clusters requires the definition of a meaningful distance measure on the data points. Automatic clustering algorithms, like k-means, or visualization tools, like the U-Matrix, use these distances. Distances alone, however, might not sufficient to describe clusters properly. Consider, for example, the TwoDiamonds data set depicted in Figure 7. The data consists of two clusters of two dimensional points. Inside each “diamond” the values for each data point were drawn independently from uniform distributions.

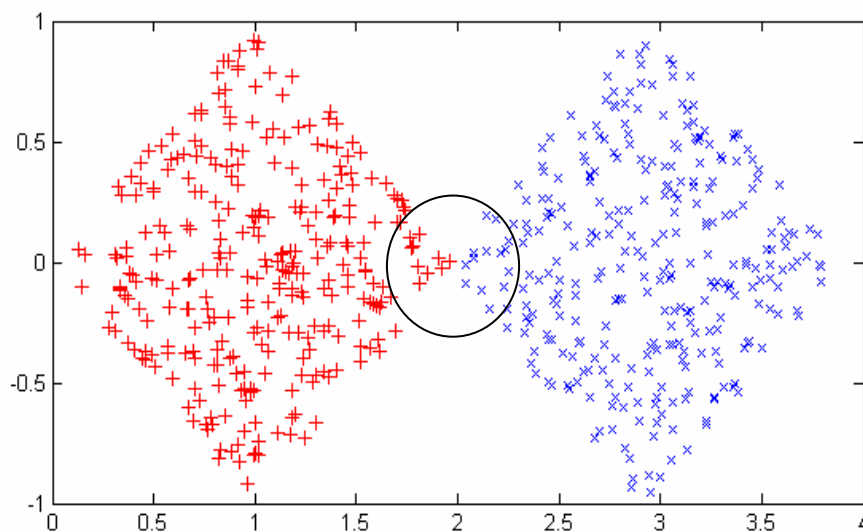


Figure 7: TwoDiamonds data set

At the central region, marked with a circle in Figure 7, the distances between the data points are very small. For distance based cluster algorithms it is hard to detect correct boundaries for the clusters. Distance oriented clustering methods such as single linkage, complete linkage, Ward etc. produce classification errors. The picture changes, however, when the data's density is regarded. The density at the touching point of the two diamonds is only half as big as the densities in the center regions of the clusters. This information may be used for the clustering of the data. Density based clustering algorithms have therefore drawn much attention in the last years within the context data mining.

A P-Matrix allows a display of the density structure of a high dimensional data set (See Figure 5). The U*-Matrix, as defined here, combines the information of local distances and local density. The calculation of the U*-Matrix is such, that inner cluster distances are depicted with lower heights than on the corresponding U-Matrix. In the extreme case, where density is maximal, U*-heights vanish.

Inter cluster distances, on the other hand, are augmented compared to the U-Matrix. This can be clearly seen by a comparison of the U- and U*-Matrix of the glioblastoma data (compare Figure 4 and Figure 6).

In the calculation of an U*-Matrix the mean and minimum of the density values of the P-Matrix are used. The validity of these values can be questioned. Using Pareto Density Estimation leads to meaningful values in the examples we have seen so far. More robust definitions, which are less sensible to outliers or skewedness in the distribution like, the median and minimum on trimmed P-heights might be also used.

In order to minimize the influence of local variations in the density estimation we recommend to apply a median filter with a 3x3 window on the P-Matrix prior to the calculation of (i). Such a nonlinear filter preserves sharp gradients in the density. It smoothes out, however, meaningless local variations in the density estimation.

The maximum value of a P-Matrix is also a canonical center point for a display of a borderless SOM, like the toroid SOM. The structures produced by the self organization of the map can be shifted anywhere on a toroid map. Taking the point with the biggest density as the center point normalizes all displays of borderless ESOM, U-Matrix, P-Matrix and U*-Matrix.

U*-Matrix heights of inter cluster distances are considerably augmented in some data sets. In some cases it has been useful to clip the resulting U*-Matrix at a maximum height of three times the maximum height of the U-Matrix.

U*-Matrix have considerably improved the visualization of structures in high dimensional data. The first U*-Matrix has been drawn at September 6th 2003 for a Data set on DNA microarrays. The analysis of this data set using U-Matrix technologies have been termed difficult elsewhere [6]. With an U*-Matrix, however, meaningful structures can be recognized [12].

Cluster analysis of the Glio data using WARD clustering showed a clear cluster structure. This could not be detected in the U-Matrix. A U*-Matrix could, however be used to cluster the data. The resulting clustering is very similar, but not identical to the WARD clustering. It has been shown by the author that the usage of ESOM leads to projections, which show a nonlinear unfolding of clusters in a high dimensional data set (Ultsch/Ringe).

It has also been demonstrated that ESOM are superior to cluster algorithms like k-Means or hierarchical cluster algorithms for some data sets. See, for example, the Chainlink data set above. Finding the “right” clustering in a data set is always difficult. Cluster algorithms of very different nature should be tried and the results compared. K-means and hierarchical cluster algorithms like WARD or “single linkage” are based on distances, while others are based on density estimation. The U*-Matrix method combines distances and densities. A version of a clustering algorithm based on gradient descent in the U*-Matrix (U*C) has been implemented and is presently being investigated [20].

8 Conclusion

A U-Matrix displays local density structures on a topology preserving projection of a high dimensional data space onto a two dimensional map. We have demonstrated in this paper that distance structures might, however, be not enough to see meaningful clusters in an U-Matrix. The combination of an U-Matrix with the recently introduced P-Matrix leads to a new visualization tool for data mining in high dimensional data sets, the U*-Matrix. This Matrix combines distance and density information. Distances inside a cluster are given less weight than distances between clusters. In a practical example it could be demonstrated that cluster structures that are detected by other clustering algorithms can be seen on an U*-Matrix

Density estimation allows a different insight into a high dimensional data set than a display of distance structures. In an U*-Matrix both information is combined. This might be very useful for data mining and knowledge discovery. It might also be used for a self organized clustering algorithm.

Acknowledgements

The author wishes to thank Dr. Iglesias-Rozas for the provision of the data and the members of the Databionics research group for proof reading the text.

References

- [1] T. Kohonen, "Self-Organized formation of topologically correct feature maps", *Biological Cybernetics*, Vol.43, pp.59-69, 1982.
- [2] A.Ultsch, "Data Mining and Knowledge Discovery with Emergent Self-Organizing Feature Maps for Multivariate Time Series", in: Oja, E., Kaski, S. (Eds.): *Kohonen Maps*, pp. 33 - 46.
- [3] A.Ultsch, "Self-Organizing Neural Networks for Visualization and Classification", *Proc. Conf. Soc. for Information and Classification*, Dortmund, April 1992.
- [4] A.Ultsch, M.Eilers, "DNA Microarrays of tumors diagnosed with databionic methods" (in German) in *Kooperationspartner in Forschung und Innovation*, pp 19 - 20, Wiesbaden, 2002
- [5] A. Ultsch, Röske, F.: *Self-Organizing Feature Maps Predicting Sea Levels* , in *Information Sciences* 144/1-4, Elsevier, pp 91 - 125, Amsterdam, 2002
- [6] S.Kaski, et al, *Analysis and Visualisation of Gene Expression Data using Self Organizing Maps*, Proc NSIP, 1999.
- [7] A.Ultsch, *Maps for the Visualization of high-dimensional Data Spaces*. Proc. Workshop on Self Organizing Maps, WSOM03, pp 225-230.
- [8] A. Ultsch, *Emergent Self-Organizing Feature Maps used for Prediction and Prevention in Mobile Phone Markets* , in *Jornal of Targeting* 10/4, Steward, pp 401 - 425, London, 2002
- [9] A. Ultsch, *Fundamentale Aktienanalyse mit selbstorganisierenden Dataminingmethoden* , in *Kooperationspartner in Forschung und Innovation*, pp 25 - 26, Wiesbaden, 2001
- [10] D.W. Scott, "Multivariate Density Estimation", Wiley-Interscience, 1992.
- [11] Ultsch, *Pareto-80/20-Regel: Eine Begründung der Pareto 80/20 Regel und Grenzwerte für die ABC Analyse* , in *Technical Report, Nr 30, Marburg*, 2001
- [12] A. Ultsch, *Optimal Density Estimation in Data containing Clusters of unknown Structure*, in *Technical Report No. 34, Department of Mathematics and Computer Science Philipps-University Marburg*, 2003
- [13] A. Ultsch, Iglesias-Rozas, J.R.: *Knowledge Discovery in Glioblastoma Data*, to appear, 2004.
- [14] A. Ultsch, *Maps for the Visualization of high-dimensional Data Spaces* , in *Proc. Workshop on Self organizing Maps*, pp 225 - 230, Kyushu, Japan, 2003.
- [15] J. Vesanto et al., "Self-organizing map in Matlab: the SOM toolbox", *Proceedings of the Matlab DSP Conference*, pp 35--40, Espoo, Finland, November, 1999
- [16] A. Ultsch, *Self Organizing Neural Networks perform different from statistical k-means clustering* In: M. van der Meer, R. Schmidt, G. Wolf, (Eds.): *BMBF Statusseminar "Künstliche Intelligenz, Neuroinformatik und Intelligente Systeme"*, pp. 433 - 443, 17. - 19. April 1996, München.
- [17] R.A. Fisher, "The use of multiple measurements in taxonomic problems", *Annals of Eugenics* 7, 179-188, 1936.
- [18] Rand, W. M. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66, 846–850, 1971.
- [19] A. Ultsch, *Pareto Density Estimation: Probablity Density Estimation for Knowledge Discovery* , in *Proc. GfKI Cottbus, Cottbus*, 2003
- [20] A. Ultsch, *A self organizing Cluster algorithm*. to appear.