

Self Organizing Neural Networks perform different from statistical k-means clustering

A. Ultsch

Department of Computer Science

University of Marburg

Hans Meerwein Str

35032 Marburg, Germany

e-mail: ultsch@informatik.uni-marburg.de

If Self Organizing Feature Maps (SOFMs) are enhanced by specially designed visualization algorithms like U-Matrix methods, they can be used for clustering. A common belief is, that the clustering abilities of this type of Artificial Neural Networks is identical or at least similar to the k-means algorithm used in statistics. In this paper we show by means of an nonlinear separable dataset, that the clustering abilities of SOFMs are quite different from the k-means algorithm. SOFMs were able to recognize clusters in a dataset where other statistical algorithms failed to produce meaningful clusters. The dataset used, called chainlink, may serve as a “benchmark” to compare clustering abilities of different algorithms.

1 Introduction

Clustering is the process of dividing a set of objects into a number of subsets, so called classes or clusters, such that the subsets are formed of objects which share common properties. Different clustering procedures have been proposed in the statistical domain of Exploratory Data Analysis, see for example [Jain/Dubes 88] for an overview. Artificial Neural Networks (ANN) of the so called “unsupervised leaning” type claim to adapt their structures to the structure inherent in a dataset. One particular ANN of this kind is the Self Organizing Feature Map (SOFM) proposed initially by Kohonen [Kohonen 82]. A common belief about SOFMs is, that their functioning is very similar or even identical to the statistical clustering procedure called k-means (see below) [Varfis/Versino93, Murthag/Hernandez94].

In this paper we show, that the clustering abilities of SOFMs are quite different from the k-means algorithm. SOFMs were able to recognize clusters in a dataset where statistical algorithms failed to produce meaningful clusters. The dataset used has very usual statistical properties like no correlation and normal distribution. It consists, however, of two clearly separated subsets which are not separable by any hyperplane. We believe, that this dataset, called chainlink might be a suitable “benchmark” in order to test different cluster algorithms.

The rest of the paper is organized as follows: Chapter 2 gives a short introduction on SOFM's usage for clustering. While SOFM by itself are not clustering procedures it is shown how computer-graphical displays of the ANN, called U-matrix methods, can be used to complete SOFMs to a clustering algorithm. Chapter 3 describes the structure of the dataset used to compare SOFMs with k-means. In the chapters 4 and 5 the results of applying the algorithms to the datasets are described.

2 Clustering with SOFMs

Kohonen's Self-Organizing Feature Maps belong to the class of ANN called unsupervised learning networks. For an overview on ANN types see for example [Ultsch 91, Ritter et al 90, Lipp 87]. SOFMs consist basically of two layers of so called units or neurons (see figure 1). The input layer consists of N neurons corresponding to the real-valued input vector of dimension N . These units are connected to a second layer of neurons U . By means of lateral connections, the neurons in U form a lattice structure of dimensionality M . Typically M is much smaller than N . Figure 1 shows a two-dimensional grid as output unit layer U .

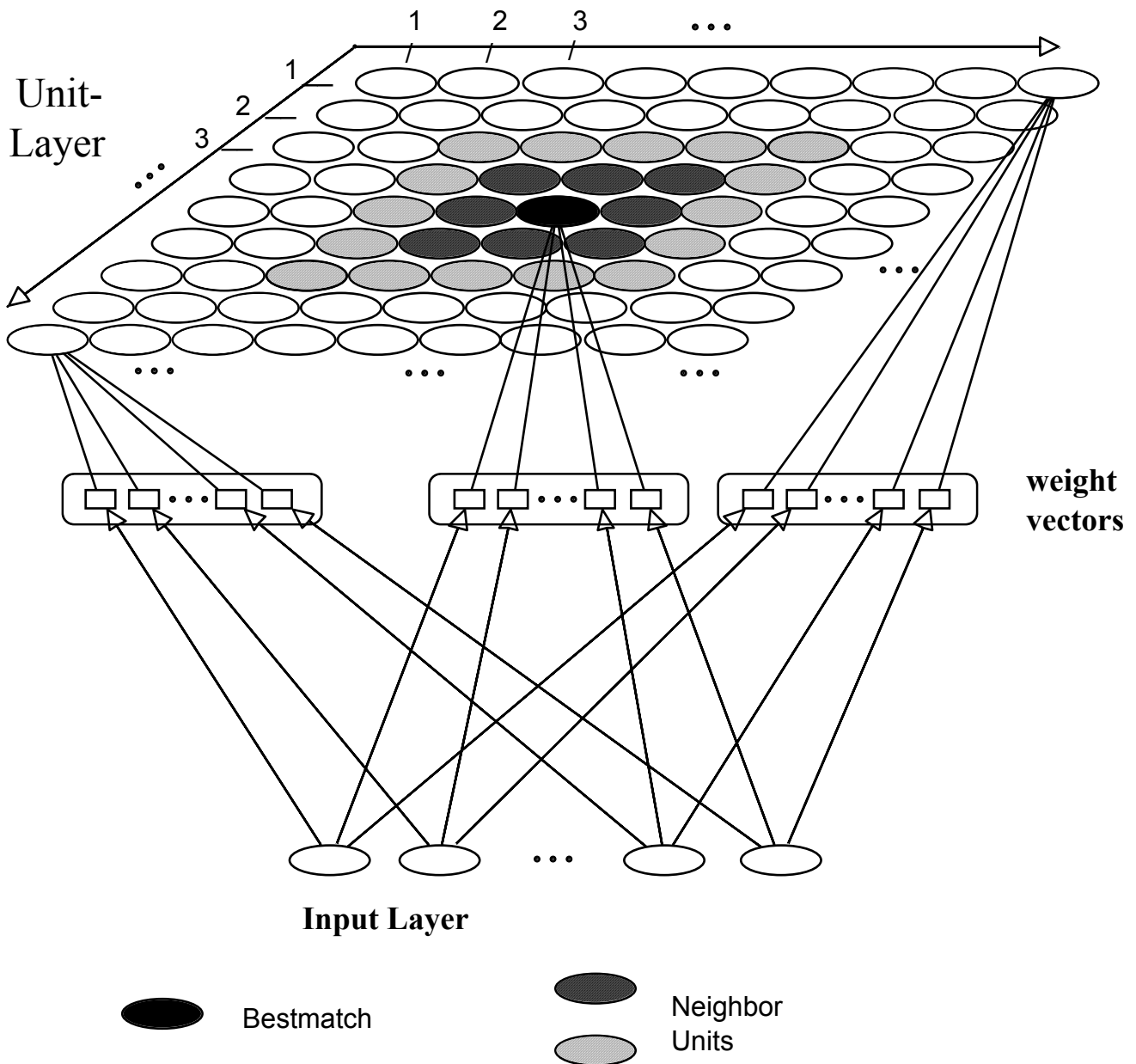


Figure 1: Structure of a SOFM

When an input data vector is presented to the network, it responds with the location of a unit in U , that corresponds most closely to the presented input. This is called the bestmatching neuron or for short the bestmatch. As a result of the learning process, i.e. the presentation of all input vectors and the adaptation of the weight vectors, the SOFM generates a mapping from the input space \mathcal{R}^n onto the lattice U with the property, that the topological relationships in input space are as good as possible preserved in U . [Ritter Schulten 86, Kohonen 84]. Similar input data should correspond to bestmatches in U that are close together. Figure 2 shows an example of an artificially generated dataset with four clusters. Four data points in the three-dimensional unit cube were chosen that constructed a tetrahedron with edge length of .96. For each of these four points ten random vectors were generated, having a length in the range from .1 to .4 with a mean length of .2. The random vectors were added to the edge points of the tetrahedron in order to get four separated data clusters.

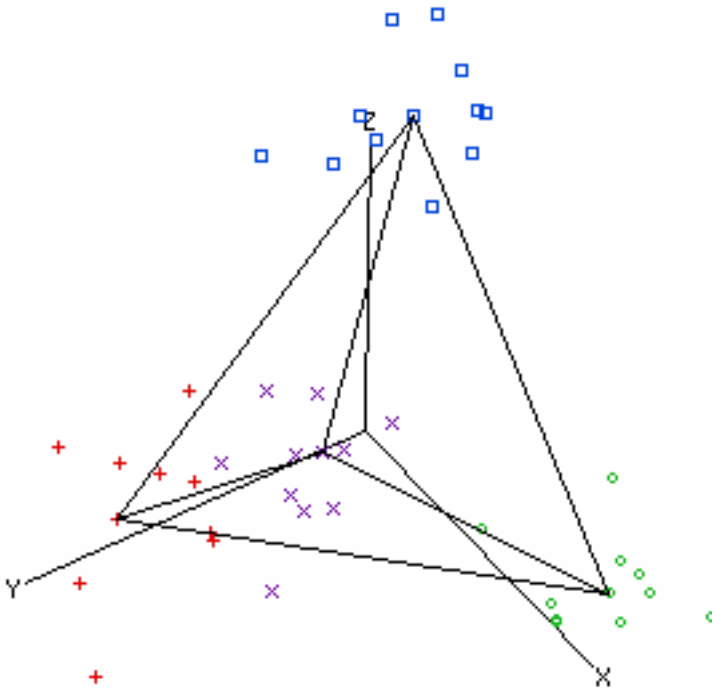


Figure 2: Exemplary dataset with four clusters

A SOFM with 3 input units and a grid lattice U with 64 by 64 units was trained with this input data. After 50.000 learning steps the bestmatches a distributed on U as figure 3 shows.

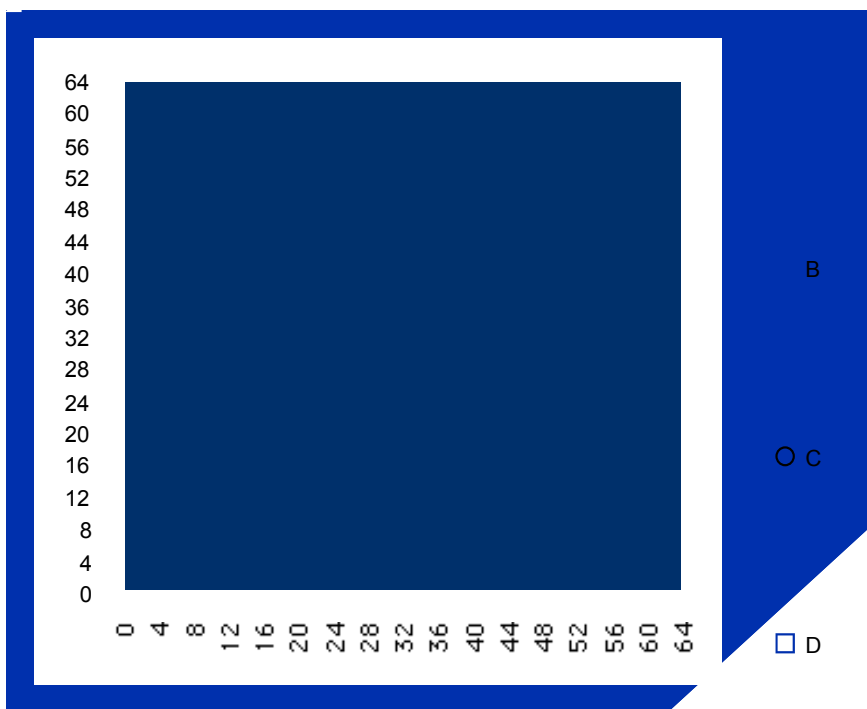


Figure 3 Bestmatches of the tetrahedron dataset on unit layer U

As it can be seen, the net arranges the clusters into the different corners of the lattice U . The distances among the data points, however, are evenly distributed. If their membership in the original clusters would not be known, *no clustering could be detected*.

In order to detect and display the structures we have developed 1990 a set of methods, called U-matrix methods (UMMs) [Ultsch 90]. The idea of UMMs is to visualize the Feature Map's topology. Analyzing the weights at each point of the grid with respect to their neighbors and then displaying the distance between two neighbor units as height, leads to an interpretable, 3-dimensional landscape of the Kohonen Map (for detail see for example [Ultsch 92]). This diagram we call Unified distance matrix or short U-matrix. The U-matrix contains a geometrical approximation of the vector distribution in the unit layer U . This display has valleys where the vectors in the lattice are close to each other and hills or walls where there are larger distances, indicating dissimilarities in the input data. A very simple U-matrix can be generated by summing up the distances (measured in input-space) in a neighborhood (in U) and displaying it on the location of a unit in U .

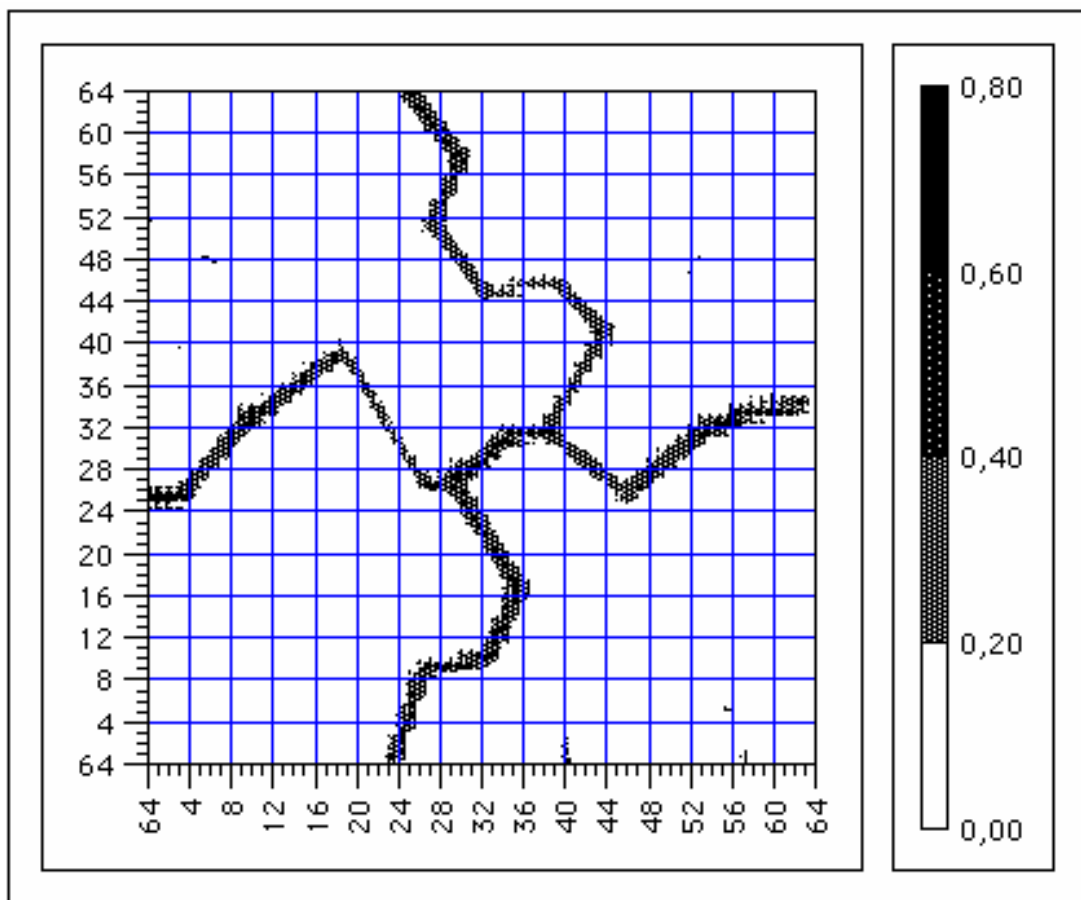


Figure 4 U-matrix of the tetrahedron dataset

To explore the properties of a n -dimensional data space we propose to interpret the landscape of the U-matrix. Given a n -dimensional data set, the question is if there exists any structure in form of subsets i.e. clusters of data that are very similar to each other. The U-matrix should be interpreted as follows: If a subset of input data falls into a valley in the U-matrix surrounded by walls then this indicates a cluster containing similar vectors. Neighbors among the U-matrix are close to each other in the \mathcal{R}^n . The dissimilarity among the different clusters is indicated by the height of the walls or hills on the U-matrix.

As can be seen in figure 4, the example dataset is divided into four different regions, with walls indicating the dissimilarity among the clusters. In order to cluster a given data set it should be noted in particular, that no previous information on the number of clusters is needed.

3 The Chainlink Data Set

In order to compare the different clustering methods, an example should be found, which could serve as a benchmark for different clustering algorithms. Such an example data set we have found to be the “chainlink”- data set presented below.

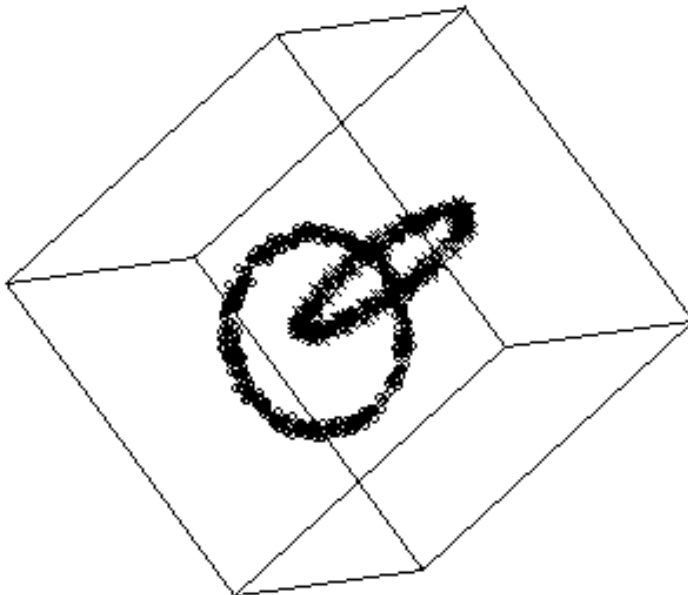


Figure 5: Chainlink Data Set

1000 datapoints in R^3 -space are arranged such, that they form the shape of two intertwined 3-dimensional rings, of whom one is extended into x-y direction and the other one into x-z direction. The two rings can be thought of as two links of a chain with each one consisting of 500 datapoints. The data is generated by a random number generator within two toroids with Radius $R = 1.0$ and $r = 0.1$ see figure 6. The data may be obtained by writing an e-mail note to the author.

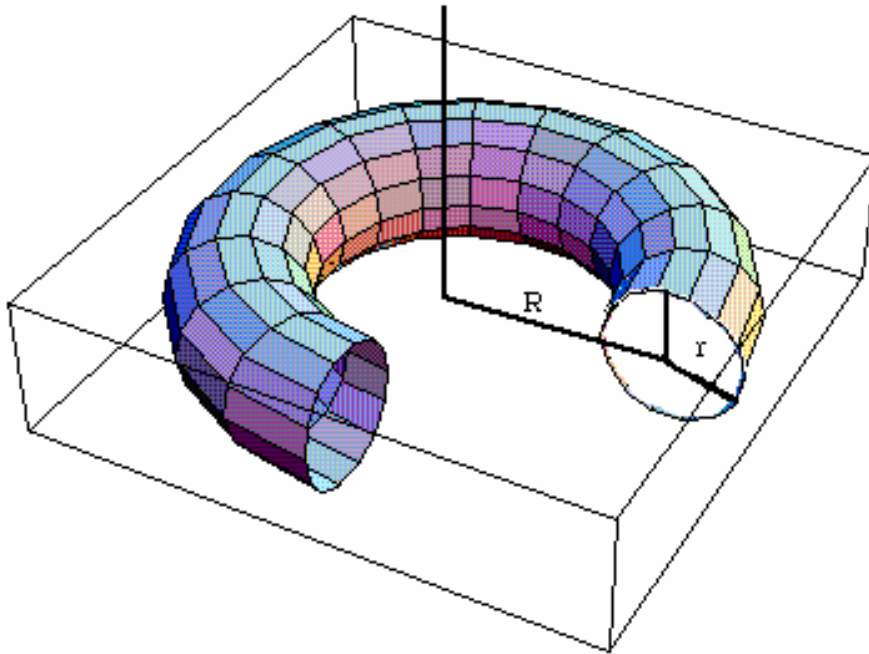


Figure 6: Toroid covering the chainlink's rings

Statistical analysis of the data shows, that the data set has the following properties

- the different components (x,y,z) of the data are uncorrelated (using Pearson's correlation coefficient)
- the distribution of each component may be regarded as *normal*
- Principal Component Analysis shows that there is no lower inherent dimensionality in the data

Details of the statistical analysis are given in [Ultsch/Vetter 94]. There exists however no (hyper-)plane that can separate the dataset correctly into the two subsets it belongs. I.e. the data set is non linearly separable. From the way the data set is created one can clearly state, however, that it consists of two clearly distinguished subsets. The innersubset distance of a datapoint is by an order of a magnitude smaller than the intersubset distance.

4 Comparison with k-means

The k-means clustering algorithm belongs to the so called "partitioning" algorithms. It was presented among others by Hartigan in 1975 [Hartigan 75] and improved in terms of speed by Hartigan and Wong in 1979 [Hartigan/Wong 79]. Variations of this algorithm were considered by many other authors, e.g. [Späth 85], [Darken/Moody90]. Its goal is it to minimize an reconstruction error, while assigning a nearest neighbor in order to compress the data vectors onto a smaller set of reference

vectors. As an inherent feature it should be noted that it is absolutely necessary to define the number of clusters as a parameter to the k-means algorithm. Seeds for new clusters are chosen by splitting on the variable with the largest distance. K-means splits a data set into the selected number of clusters by maximizing between -relative to within - cluster variation [Jain/Dubes 88]. K-means is an iterative procedure assigning the different data vectors to the specified number of non-overlapping clusters.

In a first trial k-means was used with the specification to produce two clusters. The data set was indeed split into two subsets, the datapoints in these, however, belonged to both rings. Figure 7 shows the resulting subsets.

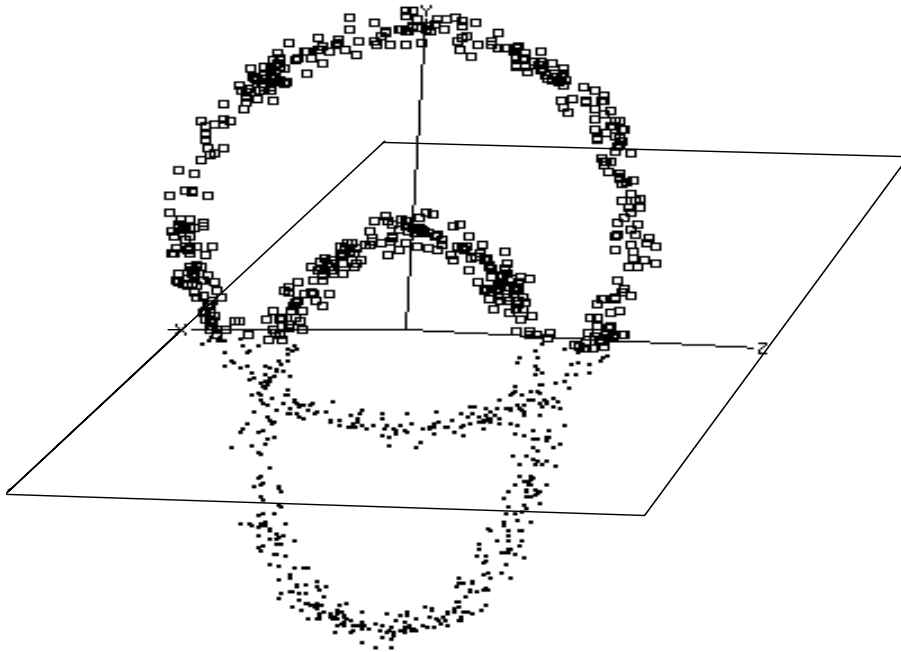


Figure 7: k-means parameterized for two clusters

One can see that k-means splits the dataset in the same way as a hyperplane cutting through the rings would do. It seems that k-means is unable to take the nonlinear structure of the data set into account.

Since the number of clusters is a parameter to the k-means algorithm, it is a common practice for exploratory data analysis to run k-means with several different numbers of clusters. K-means with four pre specified clusters produces the result given in figure 8:

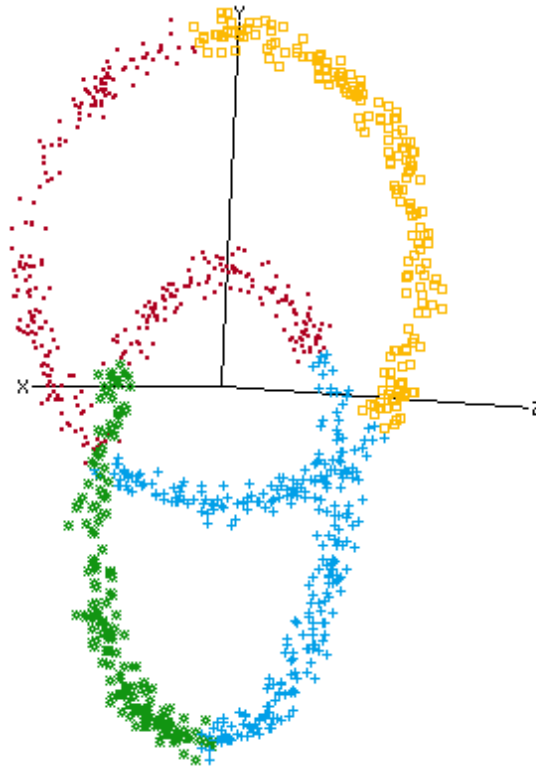


Figure 8: k-means parameterized to 4 clusters

While two of the resulting clusters contain only data of one ring, the other two, marked with a dot respectively a cross in figure 8, contain a mixture of the two rings.

In summary it can be said, the k-means algorithm is unable to classify this dataset correctly.

5 Classifying with the U-matrix technique

The chainlink dataset was used to train a SOFM with tree units in the input layer and $64 \times 64 = 4096$ Units arranged in a quadratic two-dimensional grid as output layer U. The U-matrix of this 64×64 layer is depicted in figure 9

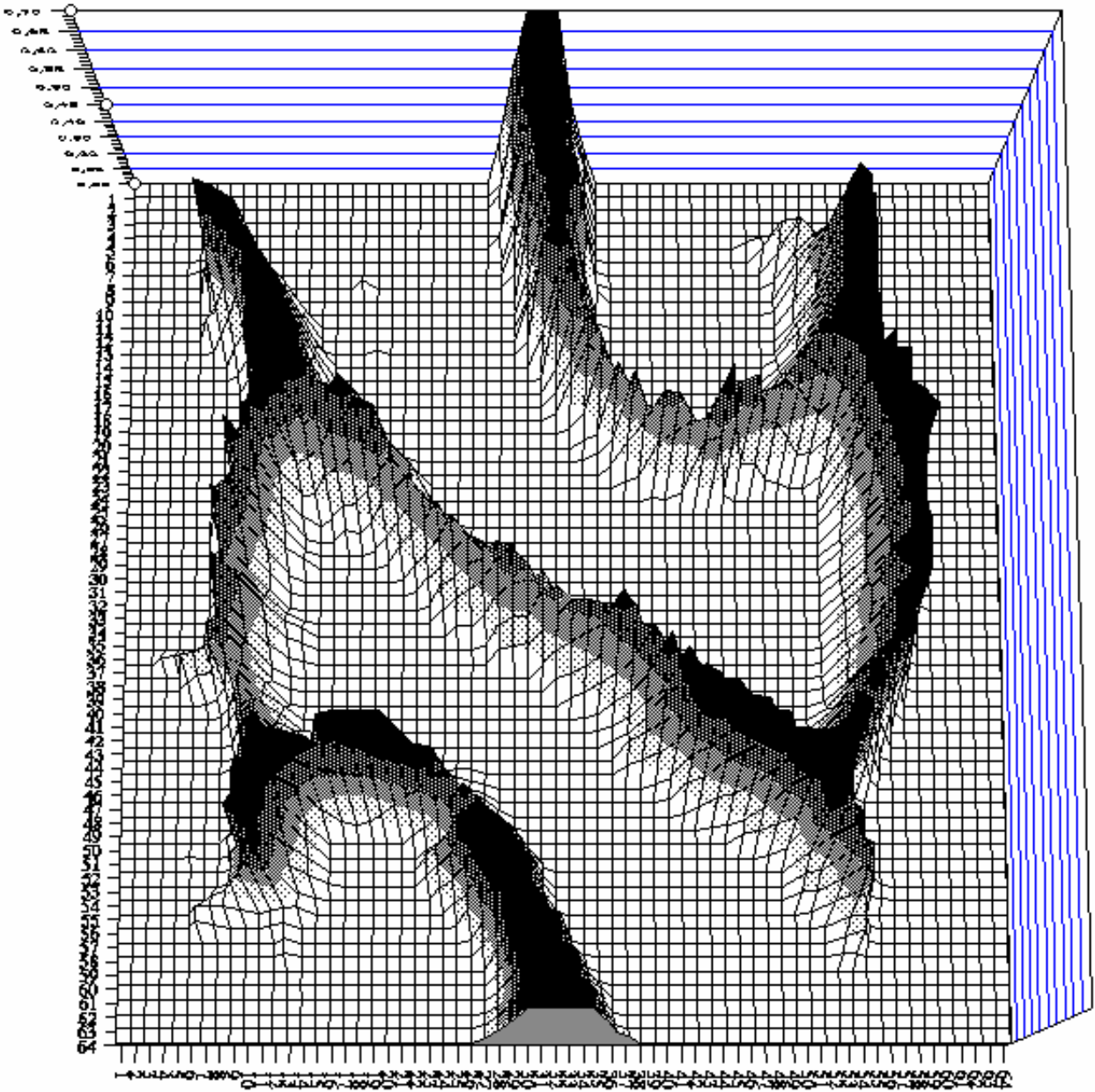


Figure 9: U-matrix of the chainlink dataset

Depicting the bestmatches of the two rings with an overlay on a top-view of the U-matrix reveals the original structure of the dataset.

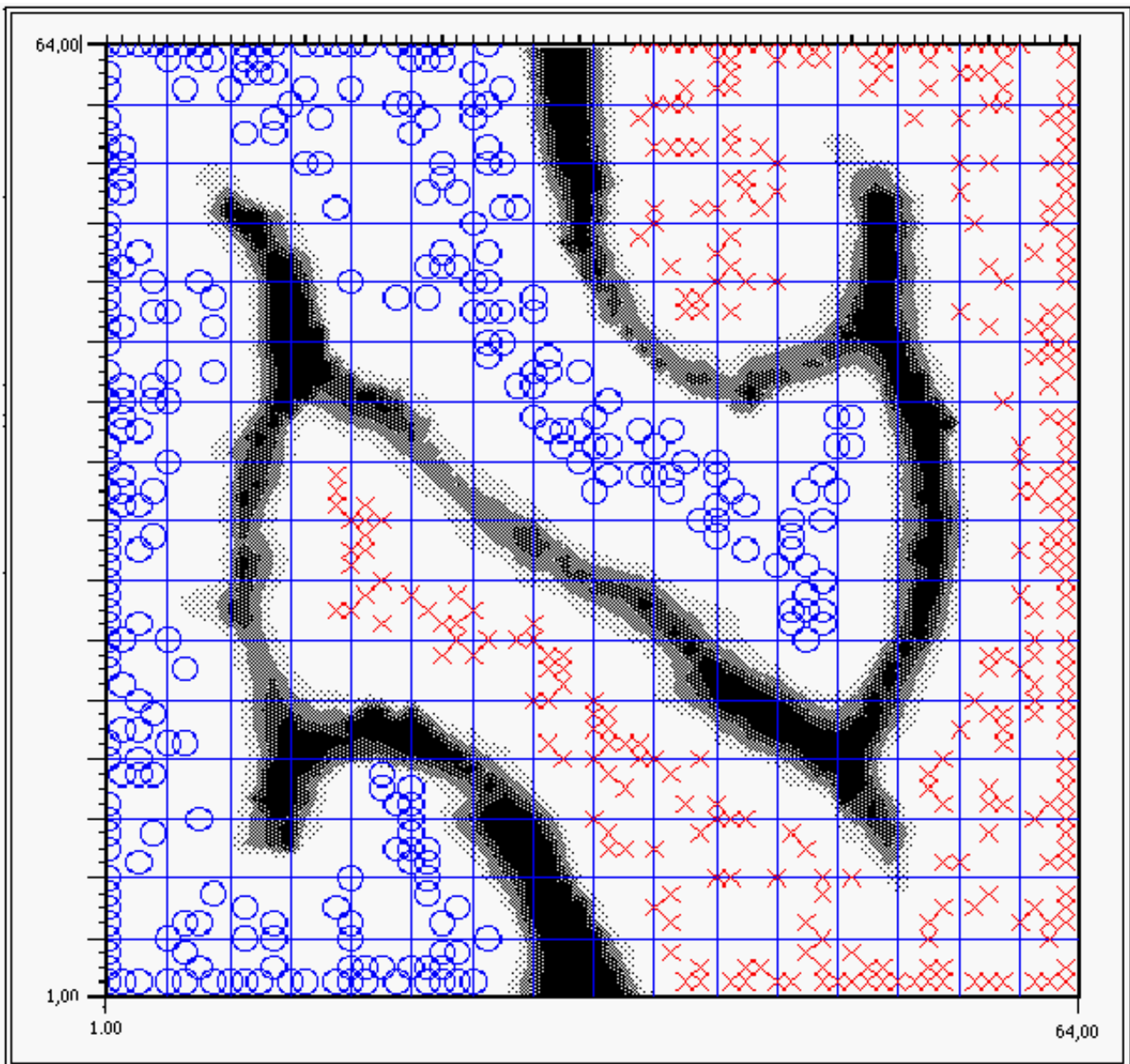


Figure 10: Bestmatches overlaid on U-Matrix for Chainlink data

One can see in figure 10 that the two rings are correctly separated on the SOFM. The areas of the subsets are intertwined like two horseshoes. Looking closer at the locations of the bestmatches one can detect, that the SOFM even tries to preserve the inter-cluster neighboring relationships.

6 Discussion

With a concise example we could show in this paper that Neural Networks of the Kohonen type exhibit, in contrast to common belief, clustering abilities, that are very different from that of the k-means algorithm. The proposed dataset has very common statistical properties like uncorrelation and normal distribution. It consists of two clearly separated subsets which are not separable by any hyperplane.

Although the inner-cluster distances are by an order of a magnitude smaller than the inter-cluster distances of the data points, k-means was unable to produce the correct clustering.

Comparing different clustering algorithms, which partially rely on different principles is a complicated task. Taking our results into account, we conclude at least, that the clustering abilities of Neural Networks of the Kohonen type are different from the k-means algorithm. Although no complete theoretical understanding of Kohonen's algorithm exist, we agree with the standpoint, that the Kohonen/U-matrix algorithm is "a form of nonlinear dimensionality reduction that has no statistical analog" [Warren 94].

We do not believe, however, that our results show a general superiority of the Kohonen/U-matrix algorithm to all other clustering algorithms. Other datasets may be found, that even show the superiority of statistical clustering algorithms. This is an open research question. For many practical applications, for example in the domain of analysis and prognosis of environmental phenomena, medicine, biotechnology industrial process control and others more [Ultsch 94] [Ultsch et al. 94a], Kohonen/U-matrix algorithms have already proven their usefulness in practice.

The simplicity and clearness of the proposed chainlink dataset, however, suggests that this set might be used to test different clustering algorithms. For long years Fishers Iris dataset has been used as sort of de-facto benchmark. [Fisher 36]. This dataset is linearly separable i.e. hyperplanes can be constructed, that are able to split the dataset into correct subsets. The chainlink dataset might complement this as a nonlinear separable but simple benchmark.

7. Conclusion

Although it imposes some difficulties to compare different clustering algorithms, which partially rely on different principles, we have shown that SOFM together with the U-Matrix method performs better than other well known statistical clustering methods. It has been capable of classifying a difficult, artificially generated dataset using unsupervised learning. In [Ultsch/Vetter 94] we have done a comparison with other, hierarchical clustering, methods. It turned out that most of these are equally unable to classify the dataset. The only algorithm, that was able to classify the data set correctly was the single-linkage algorithm. This algorithm is known, however, to perform poor on "butterfly shaped" datasets. It exhibits a tendency to produce long chains of "clusters" in the data that have no meaningful interpretation. All other hierarchical clustering algorithms presented a completely misleading picture of the structure of the data.

Literature

- [Fisher 36] Fisher, R.A.: The use of multiple measurements in taxonomic problems, *Analns of Eugenics* 7, pp 36-179 1936

- [Gordon94] A.D.Gordon: Identifying Genuine Clusters in a Classification, Computational Statistics & Data Analysis 18, pp. 561-581, 1994.
- [Jain/Dubes 88]]Jain,A.K.,Dubes, R.C.: Algorithms for ClusteringData Prentice-Hall Englewood Cliffs, N.J., 1988.
- [Lipp 87] Lippmann, R.P.: An Introduction to Computing with Neural Nets, IEEE ASSP Magazine, April 1987, pp. 4-22.
- [Ritter et al. 90] Ritter,H.,Martinez,T., Schulten,K.: Neuronale Netze, Addison Wesley, Urbana, 1990
- [Ultsch 90] Ultsch, A, Siemon H.P.: Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis, Proc. Intern. Neural Networks, Kluwer Academic Press, Paris, 1990, pp. 305 - 308
- [Ultsch 91] Ultsch, A.: Konnectionist models and their integration with knowledge based systems, research report nr. 396, department of Computer Science, University of Dortmund, February, 1991 (in german)
- [Ultsch 92] Ultsch, A.: Self-Organizing Neural Networks for Visualization and Classification, Proc. Conf. Soc. for Information and Classification, Dortmund, April 1992.
- [Ultsch 94] Ultsch, A.: Einsatzmöglichkeiten von Neuronalen Netzen im Umweltbereich, in Page et al.(Eds.): Umweltinformatik HdI 13.3, ISDN 3486227238 Oldenbourg Verlag München 1994 pp. 201-226.
- [Ultsch/Vetter 94] Ultsch, A.,Vetter, C.: Self-Organizing-Feature-Maps versus Statistical Clustering Methods:A Benchmark, Research Report No 90194, Department of Computer Science, September 1994, University of Marburg.
- [Varfis/Versino92] A.Varfis, C.Versino: Clustering of Socio-Economic Data with Kohonen Maps, IDG VSP, Ispra, Italy, pp. 813-833, 1992.
- [Warren 94] Warren, S.S.: Neural Networks and Statistical Models, Proc. SAS Users Group Conference, April 1994
- [[Ultsch et al. 94a] Ultsch, A., Guimaraes, G., Halmans, G.: Self Organizing Neural Networks and hailstorm prediction, Gesellschaft für Klassifikation e.V., Oldenburg, march 1994