

Self Organized Swarms for cluster preserving Projections of high-dimensional Data

Alfred, Ultsch, Lutz Herrmann
Dezember 2009

Databionics Research Group,
University of Marburg, Germany
Hans Meerwein Str. 22
35032 Marburg
ultsch@informatik.uni-marburg.de

Abstract: A new approach for topographic mapping, called Swarm-Organized Projection (SOP) is presented. SOP has been inspired by swarm intelligence methods for clustering and is similar to Curvilinear Component Analysis (CCA) and SOM. In contrast to the latter the choice of critical parameters is substituted by self-organization. On several crucial benchmark data sets it is demonstrated that SOP outperforms many other projection methods. SOP produces coherent clusters even for complex entangled high dimensional cluster structures. For a nontrivial dataset on protein DNA sequence Multi Dimensional Scaling (MDS) and CCA fail to represent the clusters in the data, although the clusters are clearly defined. With SOP the correct clusters in the data could be easily detected.

Keywords: Swarm Organization, Projection, Visualization, Clustering, Self Organized Feature Maps, ESOM

1 Introduction

In order to grasp cluster structures in high dimensional data, it is a common approach to project the data onto a low dimensional space such that it can be visualized [Izenman 08]. Ideally such a projection produces a map in the sense of a (geo-) graphical representation of an unknown landscape. Essential requirements for these projections are that cluster structures in the high dimensional space are not hidden nor clusters mixed, and that structures are not artificially added by the projection method. Cluster structures that are very clearly defined by a strong similarity between the cluster members and large dissimilarities between different clusters may be entangled in high dimensional space. To preserve all topological relations is, however, in principle impossible, if the dimension of the map space is strictly lower than the dimension the data's space [Bauer et al 99].

Linear projections, such as Principal or Independent Component Analysis (PCA/ICA), are in principle unable to preserve cluster structures on linear non-separable manifolds. Non-linear projection methods like Multi Dimensional Scaling (MDS) and Sammons mapping are also unable to unfold such clusters. Self Organizing Maps (SOM) [Kohonen 97] and Curvilinear Component Analysis (CCA) [Demartines/Hérault 97] are projection methods that have demonstrated to disentangle clusters and deliver coherent cluster representations. Both algorithms try to learn topographical correct mappings by the preservation of distance relations within a certain neighborhood. This neighborhood is usually parameterized by a neighborhood radius. During the iterative construction of the mapping, i.e. unsupervised learning, this

neighborhood radius is reduced. An improper choice of an annealing scheme for this radius, however, leads to a faulty representation of topology and cluster structures in CCA and SOM.

In this paper we present a new approach for topographic mapping, called Swarm-Organized Projection (SOP). SOP has been inspired by swarm intelligence methods for clustering. SOP combines ideas from the Databot approach [Ultsch 00] with Schelling's model of racial segregation in urban neighborhoods [Schelling 69]. In classical ant based clustering algorithms swarm agents pick up and drop data. See [Bonabeau 99 et al] for an overview. In contrast to this, Databots are identified with and characterized by a particular data point [Ultsch 00]. The biological equivalent would be an individual scent or pheromone. Databots are able to move on a two-dimensional discrete grid which is finite but unbound, i.e. toroid. The dynamics of a Databot swarm is controlled by programs for walking and the appreciation (sensing) of other Databots. Cluster formation in this model depends critically on a suitable annealing scheme for the range of sensors and/or the movement distances of Databots [Ultsch 00].

In Schelling's segregation model [Schelling 69] there are two types of agents (inhabitants), blacks and whites, which reside on a two dimensional grid. Inhabitants have a more or less pronounced tolerance towards neighbors of the opposite color. An agent with an unacceptable number of opposite-color neighbours, is allowed to jump randomly to a free grid space. Schelling and others were able to demonstrate that even the smallest preferences in neighborhoods leads to a fixpoint, which is the complete demixing of the two populations (segregation) [Schelling 69], [Vinkovic/Kirman 06].

The problem of finding an annealing scheme that matches the data's structures during the construction process of the projection is addressed in SOP by a Schelling like self adaptation and demixing of clusters until a sufficient level of topographic ordering is reached. Therefore no crucial parameterization for the construction of SOPs is necessary. On critical benchmark data it is demonstrated that SOP significantly outperforms SOM and CCA. For a nontrivial dataset of Bioinformatics, it is shown that MDS and CCA fail to represent the clearly defined cluster structure. With SOP the correct classes in the data could be easily found.

2 Swarm Organized Projection (SOP)

A new algorithm for the projection of high-dimensional data onto two dimensional maps grids is defined. It is called Swarm-Organized Projection (SOP). As in the Databots model [Ultsch 00], Swarm-Organized Projection agents are identified with input samples. Each agent resides on a large but finite two dimensional discrete map space grid $O \subset \mathbb{N}^2$ embedded on the surface of a torus [Ultsch 03]. Let $D = \{x_1, \dots, x_n\}$ denote the dataset with $\|\cdot\|_D$ as distances between the data in D and $\|\cdot\|_O$ denote the distances between position on the map space. The position of an agent representing data set x is called $m(x)$. At each position $i \in O$ an agent carrying a data point x can calculate its stress $\Phi(x, i): D \times O \rightarrow \mathbb{R}_0^+$

$$\Theta(x, i) = \frac{\sum_{y \in D} (h_\sigma(\|m(y) - i\|_O) \cdot \|x - y\|_D)}{\sum_{y \in D} h_\sigma(\|m(y) - i\|_O)}$$

The stress $\Phi(x, i)$ depends on a neighborhood function h_σ . This neighbourhood function $h_\sigma: \mathbb{R} \rightarrow [0, 1]$ is a symmetric and monotonic decreasing function of the map space distance from position i having it's maximum value at distance zero, i.e. at position i . An example of such a

neighbourhood function is a two dimensional Gaussian bell with variance σ^2 . The neighbourhood radius σ is adjusted (reduced, annealed) during the construction of a SOP mapping. An agent may move from a node i to a different node j , if by this move the stress at $\Phi(x, j)$ is decreased. Candidates for the target position j of a move are randomly drawn from the set of all positions such that the probability of the drawing follows a two dimensional iid normal distribution $N^2(i, s)$ centered at i with variance s^2 . A SOP mapping m is constructed by the movement of the agents until convergence is reached at a certain radius σ , then the radius is reduced. The construction of a SOP mapping can be formulated in pseudo-code as follows:

```
1: function learning of  $m = \text{SOP}(D)$ 
2: for all  $x$  in  $D$ : assign an initial random position  $m(x)$  on the grid  $O$ 
3: for  $\sigma = \{s_{\max}, s_{\max}-1, \dots, 1\}$  do
4:   repeat
5:     for all  $x$  in  $D$ :  $m(x) := \arg \min_j \Phi(x, j)$  with  $j$  drawn by  $N^2(i, \sigma)$ 
6:   until  $m$  fix
7:   return  $m$ 
8: end function  $\text{SOP}$ 
```

s_{\max} denotes the maximal distance between any two positions on the output grid O .

First, the agents are randomly located on the grid (line 2). Then, for each radius σ a fixed point iteration with respect to mapping m is performed (lines 3 to 6). Agents move simultaneously, iff they can decrease their personal amount of topographic stress (line 5). At each level of a neighbourhood radius the learning proceeds until no agents wants to move (line 6). I.e. no agent is able to find a position, where its stress is lower than at the current position.

The number of agents moving at a radius σ may vary, depending on the structure of the particular dataset. This adaptive mechanism discards the need for a pre-defined annealing scheme for σ . This means a self adaptation of the number of iterations for each value of σ .

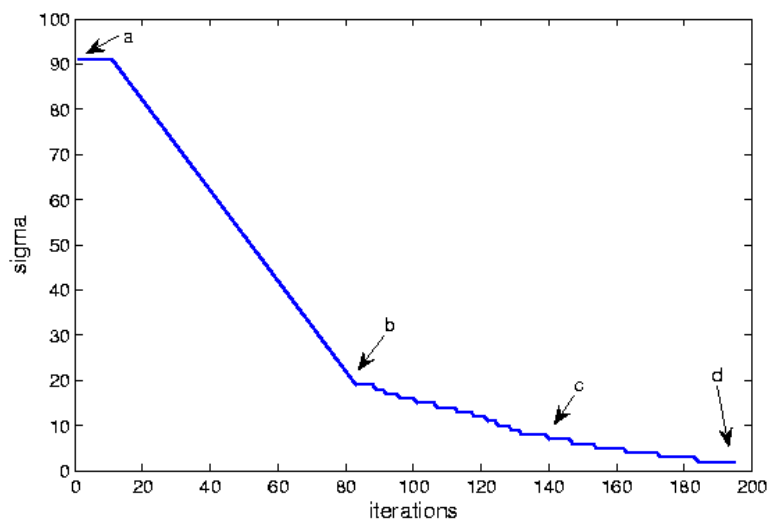


Figure 1: neighbourhood radius as a function of iterations, SOP maps Chainlink onto a 50 x 82 grid

Figure 1 shows the number of iterations that SOP used for each σ on the Chainlink data (see below). Initially (0-a), when the mapping is random, many iterations are necessary to reach convergence. So neighborhood radius σ is kept constant. When a raw topographic mapping is achieved (a), convergence is reached much faster and the radius decreases fast (a-b). This depends on the scaling of the structural features of the data. When the global (cluster-specific) features are represented a fine grained optimization within the clusters takes place (c-d). In this phase all the map space is covered by the SOP agents. As can be seen the distance structure of the data set itself determines the annealing of the radius σ . For SOP no prior knowledge of the structural features of the data is necessary to determine a suiting annealing scheme for the neighborhood radius σ . Note that SOP does not require the data as high dimensional points in a vector space. SOP's input is only the distances $\| \cdot \|_D$ between the data in D . So SOP may also be used in cases where only distances are known. Below a real world situation of this type is presented, where only distances between genetic codes are given.

3 Benchmarking Swarm Organized Projections

SOP is tested here on a number of datasets that are published in order to assess the performance of clustering algorithms (<http://www.uni-marburg.de/fb12/datenbionik>). This repository, called Fundamental Clustering Problems Suite (FCPS) contains a set of benchmark problems that test the limits of clustering algorithms. All data sets from FCPS with a dimension greater than two are used. The datasets used are as depicted in figure 2.

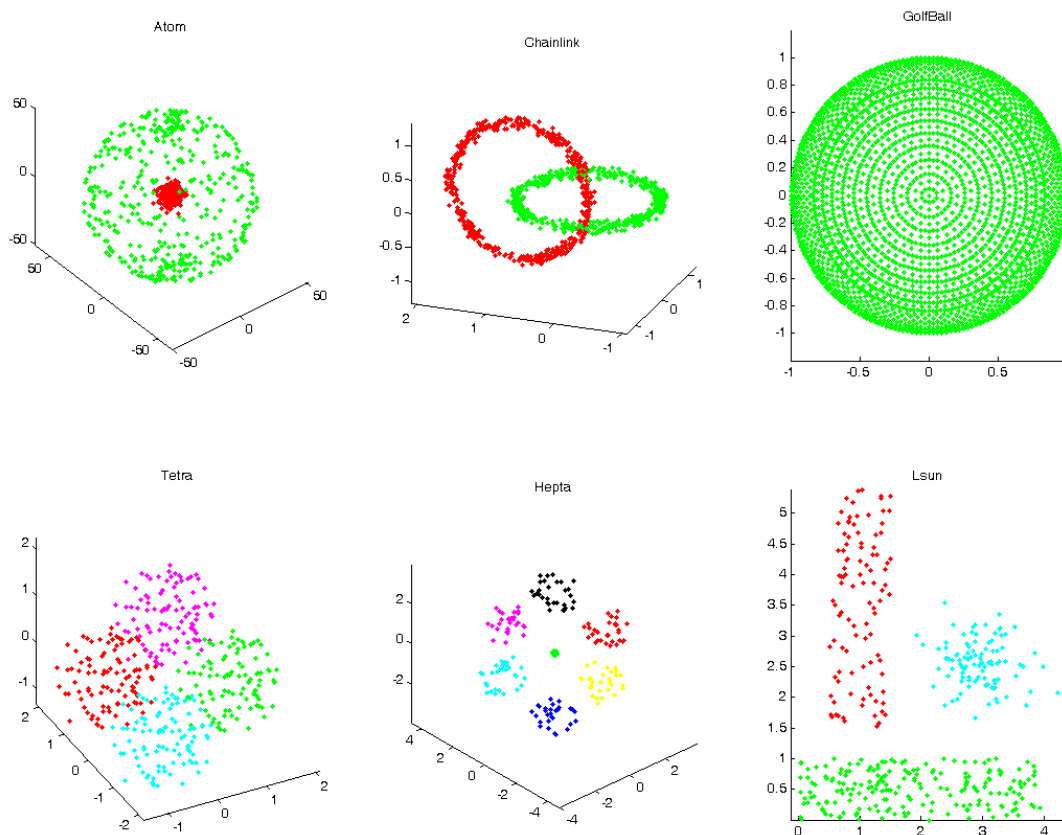


Figure 2 The Benchmark data sets from FCPS (<http://www.uni-marburg.de/fb12/datenbionik>)

The clusters in Atom and Chainlink are not separable by a linear manifold. In Atom one of the two clusters (core) is much more dense than the other (hull). The GolfBall data does not contain any identifiable clusters. All points sit on a sphere like the dimples on a golf ball. Tetra and Hepta have compact clusters which are linear separable from each other. In Tetra the clusters almost touch, i.e. the minimal inter cluster distance is small compared to the inner cluster distances. In Hepta one of the clusters is very dense. Iris data is used to show the effects of different principal axes in the clusters. The performance of SOP is tested in comparison to other projection algorithms from data spaces onto a two dimensional output. The tested projections are Principal Component Analysis (PCA), Independent Component Analysis (ICA), Multidimensional Scaling (MDS), emergent SOM (ESOM) and Curvilinear Component Analysis (CCA). We assume the reader to be familiar with PCA, ICA and MDS. For more information on these projections see, for example, [Duda et al 2003] or [Izenman 08].

The Self-Organizing Map (SOM) is a unsupervised learning artificial neural network [Kohonen 97]. SOMs use low-dimensional regular and finite grids as map space. Elements of map space (grid nodes) are called neurons or units. Each neuron i is associated with a so-called codebook vector w_i in D . Input samples are projected onto the bestmatching unit (BMU), i.e. $bm_u(x) = m(x) = \arg \min_{i \in O} \|x - w_i\|_D$. SOM is very similar to k-means clustering, if the number of neurons is small [Utsch 03]. Here, we use the SOM as a topographic projection. Therefore, we assume that the map space is sufficiently large, such that each input data can in principle be mapped onto a separate neuron and there is some empty space between the BMUs, i.e. there exist neurons that are not BMUs of input samples. Using U-Matrix methods ([Utsch 03]) on this type of SOM leads to the emergence of structural features in the data [Utsch 07]. Therefore, this type of SOM is called Emergent SOM (ESOM) as opposed to k-means-SOM. Output map grids of size $50 \times 80 = 4000$ neurons on a toroid grid were used for the experiments reported here.

Curvilinear Component Analysis (CCA) is loosely inspired by MDS and SOM [Demartines/Hérault 97]. A CCA is learned by performing for all data x, y a stochastic gradient descend on $E(x,y)$:

$$E(x, y) = \left(\|x - y\|_D - \left(\|m(x) - m(y)\|_o \right) \right)^2 \cdot h_\sigma \left(\|m(x) - m(y)\|_o \right)$$

The user has to specify a suiting annealing scheme for σ . During construction of a CCA mapping the projected position of a datapoint $m(x)$ is temporarily fixed, and all other $m(y)$ move in order to adjust the pairwise distances. Convergence in CCA is enforced through a decreasing learning rate applied during the gradient descent [Demartines/Hérault 97].

A central claim for a mapping used for the visualization of high dimensional data is the preservation of cluster structures in the data. Therefore we use an assessment measure which measures cluster preservation. This measure is called Dispersion [Herrmann/Utsch 10]. Dispersion quantifies the (dis-) connectedness of cluster-wise Delaunay graphs on the map space with respect to input distances. A cluster $C \subset D$ is cohesively mapped onto map space iff the smallest subgraph of MD containing the pictures of the points in C is connected. A topographic mapping is called cohesive iff all cluster C_1, \dots, C_k are cohesively mapped onto the map space. For all non cohesive mappings of a cluster, the dispersion of a cluster C is measured as the sum of the input distances added up along a minimal path in output space. Let $w_C(x, x')$ be class-sensitive weight function defined as follows:

$$w_C(x, x') = \begin{cases} 0 & \text{if } \{x, x'\} \subseteq C \\ \|x - x'\|_D & \text{otherwise} \end{cases}$$

The Dispersion $\text{disp}(C)$ of class C is quantified using the minimum spanning tree (MST) on MD with edge weights $w_C(x, x')$. Let $\rightarrow_C \subset D \times D$ denote the parental relation on MST with $x \rightarrow_C x'$ iff x is parental node to x' in MST. An ancestor relation \rightarrow_C^* is then obtained by $x \rightarrow_C^* x'$ iff $\exists n \in \mathbb{N}$ and there are x_1, \dots, x_n with $x \rightarrow_C x_1 \rightarrow_C \dots x_n \rightarrow_C x'$. The Dispersion of cluster C is then defined as:

$$\text{disp}(C) = \begin{cases} w_C(x, x') & \text{if for } x \in C \exists x', y \in C : x' \rightarrow_C^* x \text{ and } x \rightarrow_C^* y \\ 0 & \text{otherwise} \end{cases}$$

Dispersion $\text{disp}(D)$ of a mapping is the sum of cluster dispersions divided by the median of inter cluster distances mid :

$$\text{disp}(D) = \frac{1}{\text{mid}} \sum_{i=1}^k \text{disp}(C_i)$$

The normalization of disp by mid accounts for the data-dependent levels of scaling, different cardinalities and data manifold structures. Dispersion $\text{disp}(D)$ adds path lengths, measured by input space distances. Inner-cluster distances of a coherent projected cluster are not accounted for. A mapping which preserves the cluster structures (cohesive topographic mapping) gives a $\text{disp}(D)$ value of zero.

4 Results of the Benchmark

Dispersion of the projection of the data sets are given in the following table.

Dataset	PCA	ICA	MDS	Sammon	CCA	SOM	SOP
Atom	0.72	2.07	0.62	0.00	0.00	1.91	0.00
Chainlink	92.44	25.08	51.60	58.76	72.38	0.00	0.00
Tetra	48.35	30.69	36.16	0.00	0.00	0.00	0.00
GolfBall	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Hepta	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Iris	0.00	7.16	0.00	0.00	0.00	1.14	0.00

Table 1: Dispersion of the projections on the benchmark data

CCA and SOM seem to have abilities to disentangle the data SOM is able to unfold Chainlink, makes, however small errors on Atom and Iris. To our experience the errors in SOM grow considerably, if bordered output grids (planar grids) are used. SOP is the only projection method that consistently projects all cluster structures in the data sets.

5 SOP for Bioinformatics Data

In this chapter SOP is applied to a dataset from Bioinformatics containing protein data with a well defined cluster structure. The data set, called GPD194, was published by Popescu et al. in [Popescu 06]. It contains 194 proteins, which belong to three distinct classes of proteins: myotubularins (MTM), receptor precursors (RET) and collagen alpha chains (COL). The data

set is given as pairwise dissimilarities. These dissimilarities are calculated from the output of the BLAST algorithm [Altschul 97] as the similarity of the genetic code of the proteins. For details on this calculation see [Popescu 06]. The GPD194 data set offers quite a variety of dissimilarities between members of a protein class: the myotubularins (MTM), are very similar, the receptor precursors have many isoforms. The collagen alpha chains (COL) are, however, quite diverse. A silhouette plot [Kaufman/Rousseeuw 90] of the GPD194 data is shown in Figure 3.

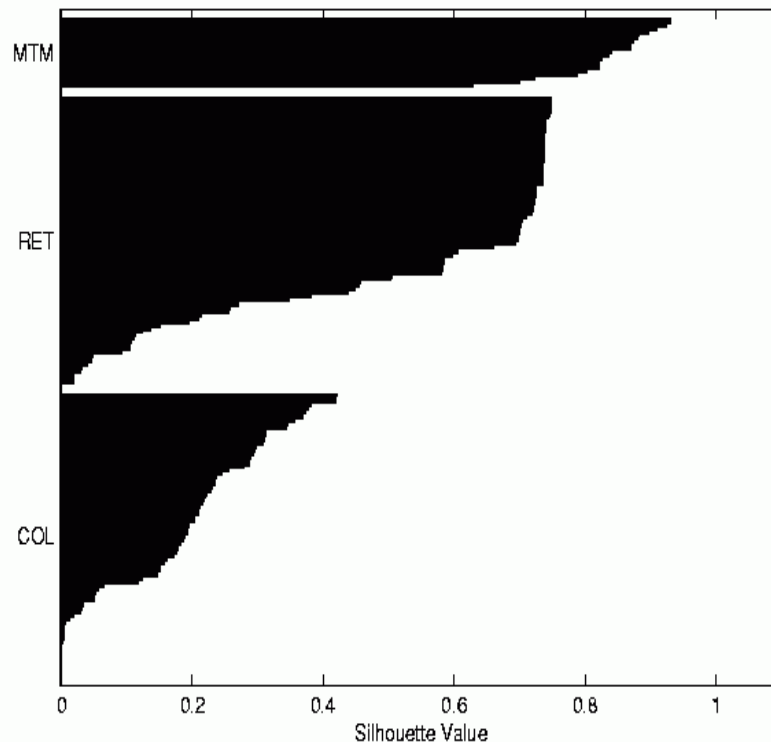


Figure 3: Silhouette plot for GPD194 data indicating three clearly separated cluster

Positive silhouette values indicate a unique and distinctive membership of an input sample to its class. The collagens (COL) cluster contains a number of proteins with silhouette values of zero. This indicates almost equal distances of these proteins towards all clusters. As can be concluded from the silhouette plot in Figure 3, the protein clusters are well-separated and a cohesive topographic mapping should be possible.

The GPD194 data set is non-vectorial: only distances are given. An embedding in Euclidean vector space produces considerable and nonvanishing errors. Therefore, SOM, ICA and PCA could not be used on this data. So only MDS and CCA could be used for a comparison with SOP.

MDS maps proteins onto a circle in map space. See Figure 4. Well-separated clusters in the input space are not indicated by MDS. For example, MTMs do not appear isolated from receptors and collagens, despite MTMs consists of a distinct set of proteins. Faithful representation of the cluster structure of GPD194 data is therefore not possible with MDS.

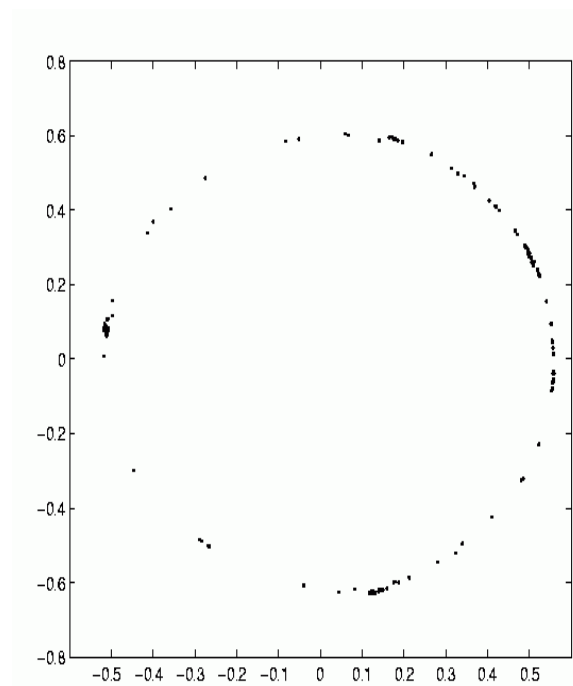


Figure 4: MDS projection of GPD194 Protein data set.

CCA depicts the proteins of GPD194 as a nearly uniform point cloud. Inter- and inner-cluster distances are not preserved, e.g. receptors and collagens do not appear as well-separated clusters with large inter-cluster distances and smaller inner-cluster distances. For illustration see Figure 5 Thus, cluster structures of GPD194 data can not be faithfully retrieved with CCA.

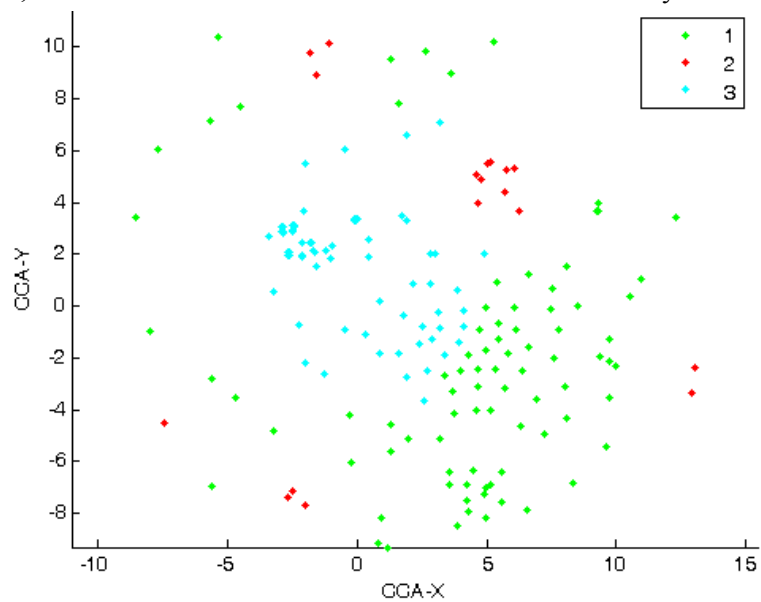


Figure 5: CCA projection of GPD194

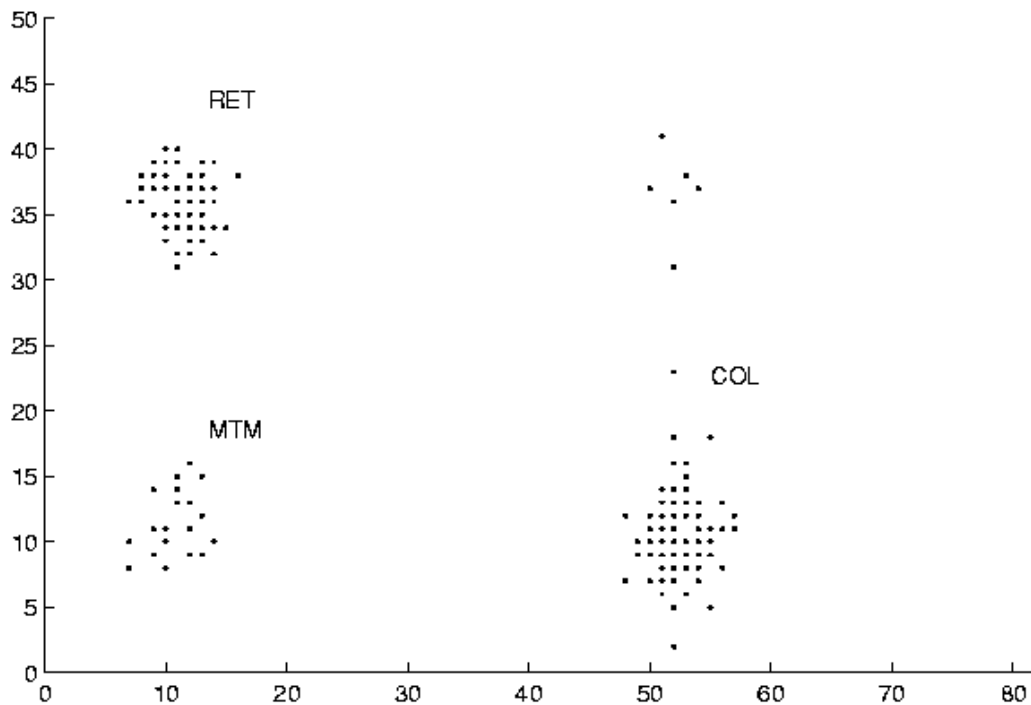


Figure 6 SOP projection of GPD194 protein distances

Figure 6 shows that SOP correctly projects the three a priori classes of proteins. In contrast to MDS and CCA, SOP depicts each class of proteins as a well-separated heap. The protein classes are easily retrieved from map space. As it was expected from the Silhouette plot, the collagens (COL) are more scattered on the grid and show some outliers.

6 Discussion

Linearly non separable data sets, of which Chainlink, Atom, and GolfBall are canonical examples, demand of a projection to unfold the structures and project the clusters in a coherent way on the output map. The coherent representation of nontrivial entangled cluster is in principle impossible for projections like PCA, ICA, MDS and Sammons mapping. This paper demonstrates that only such projection methods that exclude more and more pairwise relations during their construction are able to visualize nontrivial cluster structures topologically correct and cluster preserving.

Such projections belong to a class, called focussing projections. CCA, SOM and SOP belong to this class of projections. PCA, ICA and MDS are non-focussing projections. Focussing algorithms first capture global (inter-cluster) proximities, later more local (intra-cluster) data structures are represented. This is an effect of the shrinking neighborhood h_σ . It can be proven that the learning algorithm of SOP is sound, complete and terminating [Herrmann/Ultsch 10]

The right choice of an annealing scheme for the neighborhood radius σ is crucial for focussing projections [Nybo et al 07]. An annealing which is not well suited to the intricacies of the data's structures leads to topographical incorrect representation of clusters. The choice of the annealing scheme is, however, in practice left to some default strategy of a particular

implementation. An optimal annealing scheme depends both on the structure of the map space (e.g. bordered vs unbound, etc.) and the manifold containing the input samples. The latter is, however, an unknown quantity. In SOP the annealing adapts itself to the data's structure by a fixed point iteration. This is not to be confused with fixed points in mathematics. With regard to probabilistic agents, fixed point refers to the state where no agent has moved. Let $p \in [0,1]$ denote the upper bound of the probability that an agent moves. The probability $P(t)$ that for $t \in \mathbb{N}$ iterations at least one agent moved follows as $P(t) = (1-(1-p)^n)^t$. Since $\lim_{t \rightarrow \infty} P(t) \rightarrow 0$ the

iteration is likely to stop after a sufficient number of iterations. Therefore, SOP's learning algorithm terminates. Furthermore it can be proven that SOP's learning algorithm is sound and complete (see [Herrmann/Ultsch 10]). SOM's leaning is neither sound nor complete. For CCA it can be shown that learning is complete and sound. Yet, CCA fails to capture the non-linear manifold of the Chainlink benchmark data (see).

Dispersion is a raw measure in comparison with other topographic quality functions, e.g. Topographic Function [Bauer et al 99]. Dispersion does not account for the inner-class topography. However, the evaluation of inner-class topography preservation contributes error terms of its own which may easily blur more severe errors of non-cohesive

The performance on the benchmark data demonstrates the superiority of the focussing algorithms over the non focussing algorithms. Among the first, SOP has the advantages of a self adaptation of the annealing scheme, a provable sound, complete and terminating leaning algorithm. On first nontrivial examples SOP has demonstrated superior performance with regard to the coherent representation of clusters.

7 Summary

A novel method to project high dimensional data onto two dimensional map spaces, called Swarm Organized Projection (SOP), is presented. SOP is derived from concepts of swarm intelligence and from emergent SOM (ESOM). It can be proven that the learning algorithm of SOP is sound, complete and terminating.

In many similar algorithms like ESOM and CCA the choice of an annealing scheme for neighborhoods is crucial for a coherent representation of nontrivial high-dimensional cluster structures. SOP solves this problem trough self adaptation to the data's structures using a swarm intelligence technique.

On crucial benchmark data it is shown that SOP outperforms classical projections and other focussing projection methods. It is also demonstrated that SOP is able to represent clusters in a high-dimensional real world data set, where other projection methods fail.

8 References

- Altschul, S.F. Gapped BLAST and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402, 1997.
- Bauer,H.-U., Herrmann,M., Villmann, T. Neural maps and topographic vector quantization. *Neural Networks*, 12(4-5),p 659–676, 1999.
- Bonabeau,E., Dorigo,M., Theraulaz,G. *Swarm intelligence: from natural to artificial systems*. Oxford University Press, Inc., New York, NY, USA, 1999.
- Demartines,P, Hérault,J. Curvilinear component analysis: a self-organizing neural network for nonlinear mapping of data sets. *IEEE Transactions on Neural Networks*, 8, p 148–154, 1997.
- Duda,R.O, Hart,P.E., Stork, D.G. *Pattern Classification*, Wiley, 2003
- Herrmann,L, Ultsch, A. *Swarm-Organized Projection for Topographic Mapping*, submitted to *Neurocomputing*, Sept 2009
- Izenman, A.J. *Modern Multivariate Statistical Techniques*, Springer, 2008
- Kaufman, L., Rousseeuw, P. J. *Finding groups in data, an introduction to cluster analysis*. Applied Probability and Statistics, New York: Wiley, 1990.
- Kohonen,T. *Self-organizing Maps*, 2nd edition, Springer-Verlag, Berlin, 1997.
- Nybo,K., Venna,J., Kaski,S. The self-organizing map as a visual neighbor retrieval method. In *Proc. 6th International Workshop on Self-Organizing Maps (WSOM)*, Bielefeld, 2007.
- Popescu,M., Keller,J., Mitchell,J. Fuzzy measures on the gene ontology for gene product similarity, *IEEE Trans Computational Biology and Bioinformatics*, 3(3), 2006.
- Schelling,T.C. Models of segregation. *The American Economic Review*, 59(2):488–493, 1969.
- Ultsch, A Clustering with Databots, In: *Proc. Int. Conf. Advances in Intelligent Systems Theory and Applications (AISTA)*, p 99-104, Canberra, 2000.
- Ultsch, A. Maps for the visualization of high-dimensional data spaces. In *Proceedings Workshop on Self-Organizing Maps (WSOM 2003)*, pages 225–230, Kyushu, Japan, 2003
- Ultsch, A Emergence in self organizing feature maps. In *Proc. 6th International Workshop on Self-Organizing Maps (WSOM)*, Bielefeld, 2007.
- Vinkovic,D., Kirman,A. A physical analogue of the Schelling mode, *Proceedings of the National Academy of Sciences*, Nr 5, 103, pp19261-19265, 2006