

Visualisation and Classification with Artificial Life

Alfred Ultsch

Department of Computer Science, Phillips-University of Marburg,
Hans-Meerwein-Str, 35032 Marburg, Germany
(E-mail: ultsch@informatik.uni-marburg.de)

Abstract. Systems that possess the ability of emergence through self-organization are an particular promising approach to Data Mining. In this paper, we describe a novel approach to emerging self organizing systems: artificial life forms, called DataBots, simulated in a computer show collective behavioural patterns that correspond to structural features in a high dimensional input space. Movement strategies for DataBots have been found and tested on a real world data set. Important structural properties could be found and visualized by the collective organisation of the artificial life forms.

1 Introduction

Systems that possess the ability of emergence through self-organization are a particularly promising approach to Data Mining. Self-organization means the ability of a biological or technical system to adapt its internal structure to structures sensed in the input of the system without external intervention. A biological example for self-organization is the organisation of swarms, e.g., bee swarms. Emergence means the ability of a system to produce a phenomenon on a new, higher level. This change of level is termed in physics "mode-" or "phasechange". It is produced by the cooperation of many elementary processes. Important technical systems that are able to show emergence are in particular laser and maser. In those technical systems billions of atoms (elementary processes) produce a coherent radiation beam (Haken (1974)).

Self-Organizing Neural Networks with emergent properties have been extensively studied by us in the past (Kohonen (1982), Ultsch et al. (1990), Ultsch (1993), Ultsch (1995), Ultsch (1998a)). In this paper we describe a novel approach to emerging self organizing systems: artificial life forms. The central idea is, that a large number artificial life forms simulated in a computer show collective behavioural patterns that correspond to structural features in a high dimensional input space.

2 UD - Universe and DataBots

A UD-Universe (Umgebungs-Dynamik-Universum) is a world in which artificial life forms, so called Data Robots (DataBots), dwell. A UD-Universe

consists of a space, called UD-Matrix, (Umgebungs-Dynamik-Matrix) which provides locations, called UNodes, where a DataBot may be at a certain moment in time. By his presence on a UD-Matrix a DataBot changes the

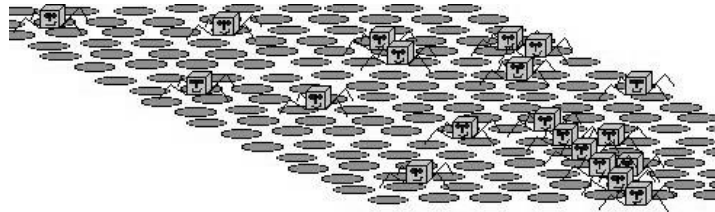


Fig. 1. Picture of a UD-UNiverse with DataBots.

UD-UNiverse. Especially a DataBot articulates its opinion (see below). A UD-Matrix is able to transmit news (opinions, scents, traces) and may be able to alter the transmissions. It may, for example, be possible to mix or weaken them. The UD-Matrix forms a constantly changing landscape for the representation of the opinions "at a glance" using U-Matrix-Methods (Ultsch (1993)). A DataBot is an artificial life form living in a UD-universe. A DataBot is able to maintain its own independent existence. By doing so it can take in food; consume food; store quantities of food (foodstuffs); stay at a certain location (UNode) in a UD-Matrix; propagate its opinion at its UNode with maximum weight and move on the UD-Matrix.

3 Movement

Grid UD-Matrices consists of neighbours in four directions. We call them North, East, South and West. The UNode itself is called O. The movement apparatus of a DataBot consist of five bins corresponding to O, N, E, S and W. These direction bins may contain positive numbers. When performing a move these numbers are rescaled to percentages. That may be regarded as probabilities for a certain direction of movement. I. e. a probabilistic choice of directions is done on basis of these percentages. If a movement is chosen by this process, all numbers in the direction bins are reset to zero. The direction of the move taken is stored in a movement memory that can be read by movement programs. Movement strategies manipulate the content in the direction bins. These programs may act simultaneously and concurring to each other. In analogy to nature we envision such movement strategies to evolve. The newer, fancier programs do not replace older ones, but work on top of them. In particular a strategy is sought for which the final location of a DataBot corresponds to the data distribution in the highdimensional vector-space of the opinions of the DataBots.

4 Movement Programs for Data Mining

A movement program is sought for which the DataBots reveal structure in a highdimensional input-dataset. The idea is to provide each DataBot with an *n-dimensional* input-vector which is the opinion of a DataBot represented in the UD-Universe. One can imagine this as a scent or smell that a DataBot emits. By sensing or smelling other DataBots, more precisely by sensing the smells that the UD-Matrix is transmitting, a DataBot searches for locations where it likes the aroma. A DataBot searches, so to speak, for UNodes where its friends are. At the same time the DataBot tries to avoid bad smells, i. e. it wants to get away from enemies. One goal of our research was to find movement programs for DataBots such that the location of a DataBot, i. e. the UNode where the DataBot wishes to stay on the UD-Matrix, reveals the structure of the highdimensional input-dataset. In particular, if there are structures like clusters in the input-dataset the DataBots should cluster too. DataBots that have data (opinions) from the same highdimensional cluster should cluster together on a UD-Matrix. They should separate themselves from other DataBots that do not belong to the same cluster. We have tried several movement programs and found in particular one of them very useful for data clustering. This movement program, called "friends_and_foes" works as follows: a DataBot gets transmitted from the UD-Matrix all smells in the neighbourhood of a certain radius. The DataBot ranks the similarity respectively dissimilarity of the highdimensional smells. The 10 % best fitting smells are considered to be from friends and the 10 % worst smelling are considered to be foes. The movement program consists of a vector addition of the direction towards the friends plus a vector addition away from the foes. The resulting direction is converted to numbers for the directional bins of the DataBot. This movement program has been successfully tested on artificial data sets containing clusters. An example of a clustering problem from real data is described in section six.

5 Softwaresystem to simulate UD-Universes

We have implemented a simulation program for UD-Matrices called DataBots. The software is written in C++ using the QT graphical library (Malorny et al. (1998)). The simulation software can display the UNodes containing the DataBots and movements. Besides that the simulation program creates a visualization of the highdimensional structure of the data using U-Matrix technologies (Ultsch (1993)). Movement strategies, which are in the focus of our present research, may be programmed and modified while a simulation of a UD-Universe is running. The movement strategies can be expressed as ASCII text using elementary operations from a DataBots functional anatomy.

6 DataBots for Data Mining

The main difference of the artificial life approach to other clustering techniques is that local movement rules for each DataBot develop a nonlinear mapping of the highdimensional data onto a two dimensional grid that take not only the closest but also the whole topology into account. The difference in performance can be seen using a data set that consists of x/y coordinates of points evenly distributed inside two tangent circles. Figure 2 shows the performance of the DataBot clustering vs. Single Linkage and Ward clustering. As can be seen errors are minimal using DataBots. In order to test the

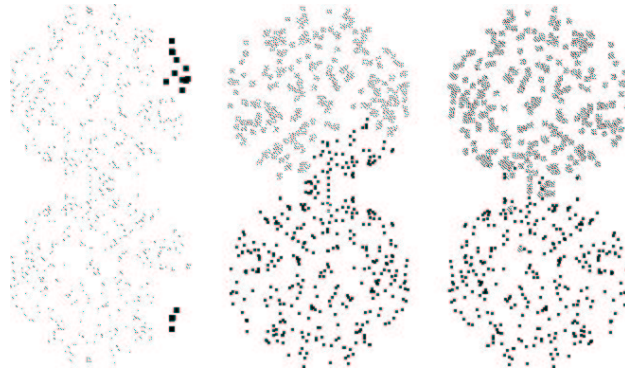


Fig. 2. Single Linkage (left), Ward (middle) and DataBot (right) Clustering.

clustering and self-organizing properties of the friends_and_foes movement strategy for DataBots on a practical example, we used a dataset made available to us by Prof. Gasteiger described in Zupan et al. (1993). This dataset has been extensively studied using statistical and pattern recognition methods, see Zupan et al. (1993) pp. 168. The dataset consists of analytical data from 572 Italian olive oils produced in nine different regions of Italy. Figure 3 contains a map of the different regions from which the olive oils are taken. For each oil the percentual contents of 8 different fatty acids are measured. I. e. the aroma of each DataBot is an 8-dimensional real-valued vector. Each DataBot was loaded with an 8 dimensional vector describing one olive oil. The number of DataBots used corresponds to the number of datavectors in the inputset. In this example we had 572 DataBots. The following picture shows the organisation of the DataBots on a 64 by 64 grid. To interpret the picture it must be understood that the picture is circular in each direction. The resulting clustering is more or less topology preserving. It can be concluded that the consistency of the olive oils vary according to the producing regions.

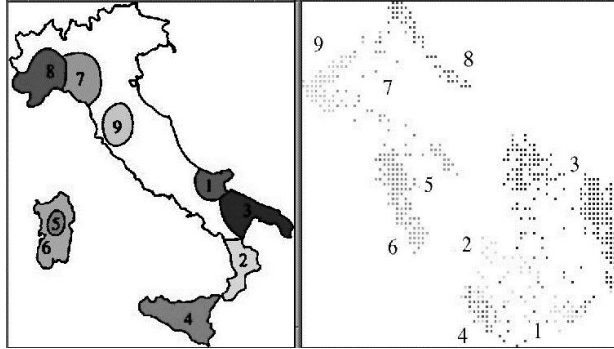


Fig. 3. Italian regions of origin of the olive oils compared to the Distribution of DataBots on Oliveoil Dataset after 400 steps.

7 Conclusion

In this work we describe a novel approach to Data Mining using artificial life forms. To our knowledge this is one of the first attempts to use artificial life forms for Data Mining and Knowledge Discovery. While definitely in its first steps the approach shows surprisingly good performance. By the usage of self-organization the system shows emergent properties, see Ultsch (1999a). It could be shown that a very simple anatomy and very simple strategies lead to suprising results concerning the detection of clusters and preserving the overall structure of highdimensional datasets.

DataBots construct a mapping from highdimensional data space onto the two dimensional grid of a UD-Universe. Thus they are similar to principal components analysis (Hotelling 1933), multidimensional scaling (Shepard 1962) and Sammon’s mapping (Sammon 1969). While principal components analysis and multidimensional scaling construct linear mappings, Sammon’s mapping and DataBots construct nonlinear mappings. The latter is in particular advantageous, if the data space is linearly non separable like in the case of the chainlink data set (Ultsch 1996). Sammon’s mapping emphasizes however the preservation of local distances while DataBots aim to preserve the overall topology. In this sense they are similar to emergent self organizing feature maps with U-matrix methods (Ultsch 1999). Our approach may lead to a very natural realization on parallel hardware using simple processors that cost less than 2 \$ in order to formulate U-Nodes and DataBots on a UD-Matrix. While the first version of DataBots was designed in April 1998 it took the work of several students mentioned below in order to provide us with a simulation tool that uses artificial life forms for the evaluation of high dimensional data structure (Ultsch 1999b).

Acknowledgements

The author wishes to thank J. Gasteiger from University of Erlangen - Nuremberg and Prof. Fiorina from University of Genoa, Italy for the olive-oil dataset. A first version of the UD-Universe has been implemented by Ingo Felger. A second version by Dirk Malorny, Ingo Müller and Falko Münchberg. The third version is currently being developed by Dirk Malorny.

References

- HAKEN, H. (1974): Synergetics, an Introduction, Springer, Berlin 1974
- HOTELLING, H. (1933): Analysis of complex statistical variables into principal components, *Journal of Educational Psychology*, 24, 417-441, 498-520, 1933.
- KOHONEN, T.(1982): Self-Organized Formation of Topologically Correct Feature Maps, *Biological Cybernetics Vol. 43*, pp 59 - 69, 1982
- MALORNY, D., MÜLLER, I. and MÜNCHBERG, F.(1998): Realization of UD-Universes, Technical Note, *Department of Computer Science, University of Marburg, Hans-Meerwein- Str., 35032 Marburg, 25. Apr. 1998*
- SAMMON, J.R. (1969): A nonlinear mapping for data structure analysis., *IEEE Transactions on Computers*, 18:401-409, 1969.
- SHEPARD, R.N. (1962): The analysis of proximities: multidimensional scaling with an unknown distance function, *Psychometrika*, 27:125-140;219-246, 1962.
- ULTSCH, A.(1993): Self-organizing Neural Networks for Visualization and Classification, in *O. Opitz, B. Lausen and R. Klar, (Eds.) Information and Classification, Berlin: Springer-Verlag, 307-313, 1993*
- ULTSCH, A.(1995): Self-Organizing Neural Networks Perform Different from Statistical k-means clustering, *Gesellschaft f. Klassifikation, Basel 8th - 10th March, 1995*
- ULTSCH, A. (1996): Self Organizing Neural Networks perform different from statistical k-means clustering, In: *M. van der Meer, R. Schmidt, G. Wolf, (Eds.): BMBF Statusseminar, pp. 433 - 443, München. April 1996, .*
- ULTSCH, A.(1998a): The Integration of Connectionist Models with Knowledge-based Systems: *Hybrid Systems, Proceedings of the 11th IEEE SMC 98 International Conference on Systems, Men and Cybernetics, 11 - 14 October 1998, San Diego*
- ULTSCH, A.(1998b): Umgebungsdynamik Universen, Technical Note, Department of Computer Science, *University of Marburg, Hans-Meerwein- Str., 35032 Marburg, 25. Apr. 1998*
- ULTSCH, A.(1999a): Data Mining and Knowledge Discovery with Self-Organizing Feature Maps for Multivariate Time Series, in *Oja, E., Kaski, S.: Kohonen Maps, p 33- 46, Elsevier, 1999.*
- ULTSCH, A.(1999b): Clustering with Data Bots, *Research Report No. 19, Department of Computer Science, University of Marburg, 1999*
- ULTSCH, A. and SIEMON, H.P.(1990): Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis, *Proc. Intern. Neural Networks, Kluwer Academic Press, Paris, 1990, pp. 305 - 308*
- ZUPAN, J. and GASTEIGER, J.(1993): Neural Networks for Chemists, *VCH, Weinheim New York, 1993*