in: Polani, D. et al.(eds): Abstracting and Synthesizing the Principles of living Systems, GWAL, Lübeck pp 191-199

Data Mining

as an Application for Artificial Life

Alfred Ultsch

Phillips-University of Marburg, Department of Computer Science

Hans-Meerwein-Str, 35032 Marburg

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

Abstract: Data Mining aims to discover so far unknown knowledge in datasets. An important step in Data Mining is the discovery of structural features in a dataset without using prior knowledge. Systems that possess the ability of emergence through self-organization are an particular promising approach to this problem. In this paper we describe a novel approach to emerging self organizing systems: autonomous minirobots, called DataBots, 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 minirobots.

1 Introduction

Data Mining aims to discover knowledge unknown so far in datasets [Ultsch 99]. An important step in Data Mining is the discovery of structural features in a dataset without using prior knowledge. Systems that possess the ability of emergence through self-organization are an particular promising approach to this problem.

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. Emergence happens in natural systems as well as in technical systems. Examples of natural emergent systems are: Cloud-streets, Brusselator, BZ-reaction, certain slime molds, a. o. m.[Haken 74]. Even crowds of human beings may produce emergent phenomena. An example is the so called „La-Ola“ wave in ballgame-stadiums. Participating human beings function as the elementary processes, who by cooperation, produce a large scale wave by rising from their places and throwing their arms up in the air. This wave can be observed on a macroscopic scale and could, for example, be described in terms of wavelength, velocity and repitition-rate. 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 [1].

in: Polani, D. et al.(eds): Abstracting and Synthesizing the Principles of living Systems, GWAL, Lübeck
pp 191-199

Self-Organizing Neural Networks with emergent properties have been extensively studied by us in the past [2], [10], [4], [5], [6], [7], [8], [9]. In this paper we describe a novel approach to emerging self organizing systems: autonomous minirobots. The central idea is, that a large number of autonomous minirobots show collective behavioural patterns that correspond to structural features in a high dimensional input space.

2 UD - Universe

A UD-Universe (Umgebungs-Dynamik-Universum) is a world in autonomous minirobots, 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.

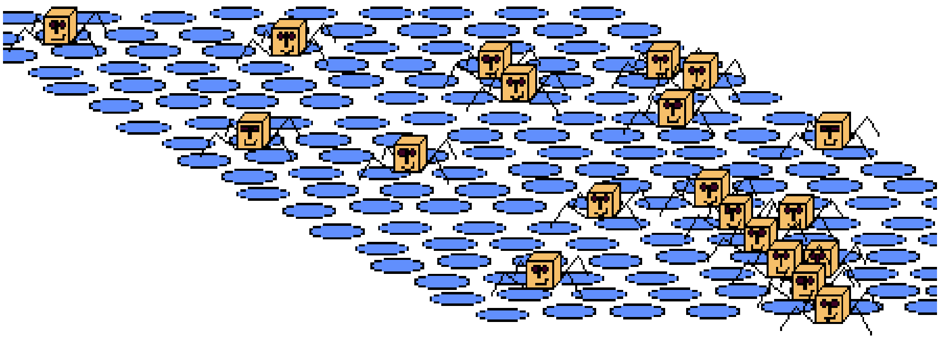

Figure 1: Picture of a UD-Universe

By his presence on a UD-Matrix a DataBot changes the 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 [4].

3 DataBots

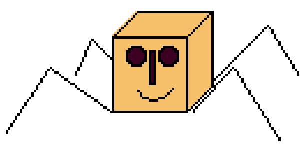

Figure 2: A DataBot

A DataBot is an autonomous minirobot roaming 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 it`s UNode with maximum weight and move on the UD-Matrix

in: Polani, D. et al.(eds): Abstracting and Synthesizing the Principles of living Systems, GWAL, Lübeck pp 191-199

DataBots need food for their existence. If a DataBot does anything or even by its mere existence its stored food (foodstuff) is reduced by food-consumption. A DataBot's foodstuff may be represented by numbers in the range of 0 % up to 100 %. If the provision is less than 100 % the DataBot may take in food from its UNode if food is provided there.

DataBots have sensors for the news (scents) broadcasted in the universe. They may store previous experiences, for example, locations where certain odors are found.

4 UD-Matrix

A UD-Matrix consists of UNodes. These are locations that are connected to their neighbours. Locations and the neighbourhood relationships might, for example, be a finite but unbounded grid of neighbouring locations.

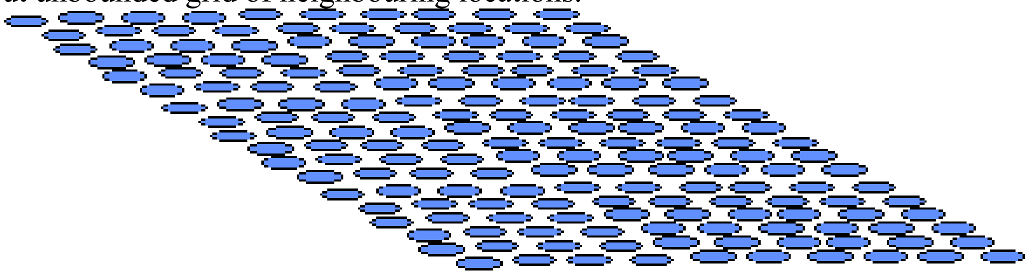

Figure 3: UD-Matrix, the ends are connected to each other

UNodes are specific locations on a UD-Matrix. They may provide food, provide a place for a DataBot, have neighbours, allow a DataBot to move to its neighbours, can be part of the visual display of a U-Matrix-Method, transmit news, may modify and transmit their own news, transmit their own news (newsbroadcast).

Two locations on a UD-Matrix are neighbours if they exchange news. The opinion of a neighbour will be transmitted to a neighbour in the opposite direction possibly in a weakened form. This means that the weight of the opinion may be diminished while transmitting. News are opinions and vice versa.

Opinions are n-dimensional vectors with attached weights. Weightless opinions, i. e. with a weight equal to zero, may disappear. Weights are simple (scalar) numbers. They range from zero (weightless) up to a maximum weight (100 %).

An opinion of a UNode may be taken from a DataBot that restsat that UNode. Otherwise it will be generated by weighting and summing up neighbouring opinions. The weight of an opinion is chosen with regard to the weights of the neighbouring opinions. The opinion of a DataBot that has its location at a particular UNode will be taken as the opinion of this UNode. The opinion represented at a UNode is propagated to all neighbours of the UNode

in: Polani, D. et al.(eds): Abstracting and Synthesizing the Principles of living Systems, GWAL, Lübeck
pp 191-199

5 Movement

Grid UD-Matrices consist of neighbours in four directions. We call them North, East, South and West. The UNode itself is called O.

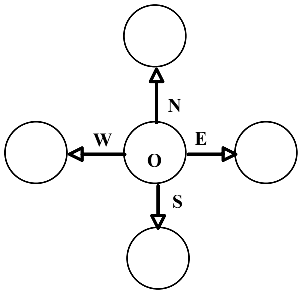

Figure 4: Neighbours of UNodes on a grid UD-Matrix

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. These are regarded as probabilities for a certain direction of movement. I. e. a probabilistic choice of directions is done on base of these percentages.

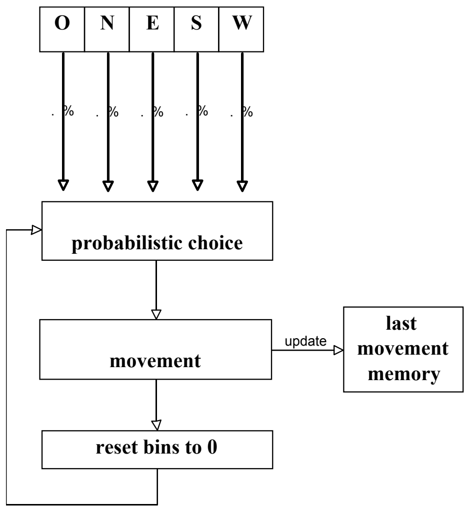

Figure 5: Functional movement anatomy of a DataBot

If a movement is chosen by this process, all numbers in the direction bins are reset to zero. The direction of the taken move 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. A very basic movement program, for example is called random (x). It adds a random number x>0 to all direction bins. The effect of this movement program is that a DataBot is randomly moving around on a UD-Matrix, with occasional stops at UNodes.

In analogy to nature we envision such movement strategies to evolve. The newer, more fancier programs do not replace older ones, but work on top of them. For example, the movement program "persistency(d)" recalls the last movement direction from memory and adds d to the bin of the recalled direction. The effect of this movement program is that a DataBot will have the tendency to continue in its chosen direction.

Other more elaborated strategies may be used and are a subject of research. In particular a strategy is sought for which the final location of a DataBot corresponds to the data distribution in the high dimensional vector-space of the opinons of the DataBots.

6 Movement Programs for Data Mining

A movement program is sought for which the DataBots reveal structure in a high dimensional 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 seperate themself 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 form the UD-Matrix all smells in the neighbourhood of a certain radius. The DataBot ranks the similarity respectively dissimilarity of the high dimensional 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.

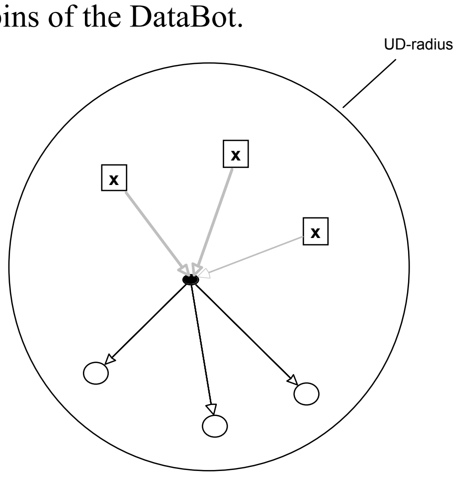

Figure 6: Friends(O) and Foes (X) in a DataBot's neighbourhood

in: Polani, D. et al.(eds): Abstracting and Synthesizing the Principles of living Systems, GWAL, Lübeck pp 191-199

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 the chapter eight.

7 Softwaresystem to simulate UD-Universes

We have implemented a simulation program for UD-Matrices called DataBots. The first version of this simulation has been written by Ingo Felger, the second called SinDBAD (= Simulation of Intelligent DataBots with Animation Display) by Dirk Malorny, Ingo Müller and Falko Münchberg. At the moment we are using the third version of the software.

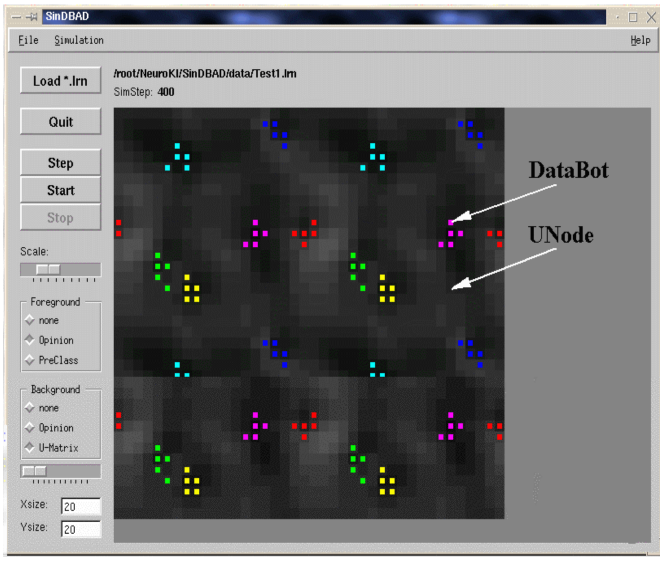

Figure 7: Screen shot of the DataBot simulation program

The software is written in C++ using the QT graphical library. As seen from the screen-shot above, the simulation software can display the UNodes containing the DataBots and movements. Besides that the simulation program creates a visualization of the high dimensional structure of the data using U-Matrix technologies [Ultsch 93]. 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 form a DataBots functional anatomy.

8 DataBots for Data Mining

In order to test the clustering and self-organizing properties of the friends_and_foes movement strategy for DataBots we took a dataset made available to us by Prof. Gasteiger described in [Zupan/Gasteiger 93]. This dataset has been extensively studied using statistical and pattern recognition methods see [Zupan/Gasteiger 93] pp 168 ff. The dataset consists of analytical data from 572 Italian

in: Polani, D. et al.(eds): Abstracting and Synthesizing the Principles of living Systems, GWAL, Lübeck
pp 191-199

olive oils produced in nine different regions of Italy. Below you can see a map of the different regions from which the olive oils are taken.

Figure 8: Italian regions of origin of the olive oils

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.

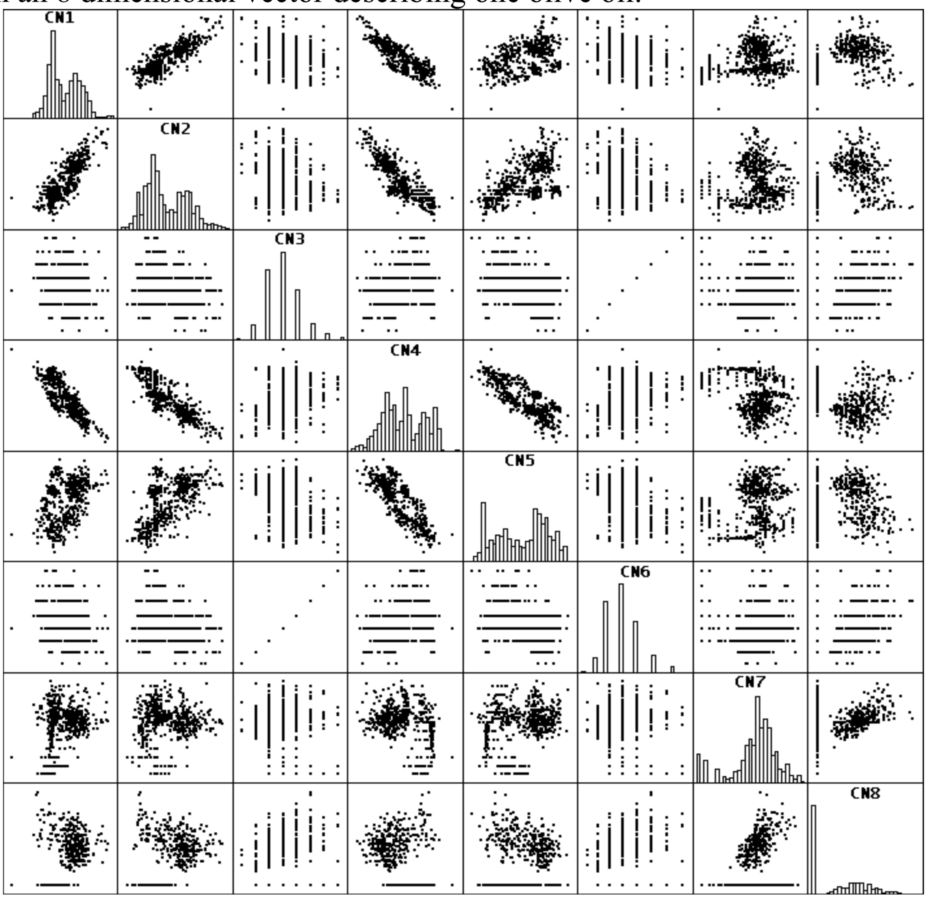

Figure 9: Scatter plot of oliveoil dataset

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.

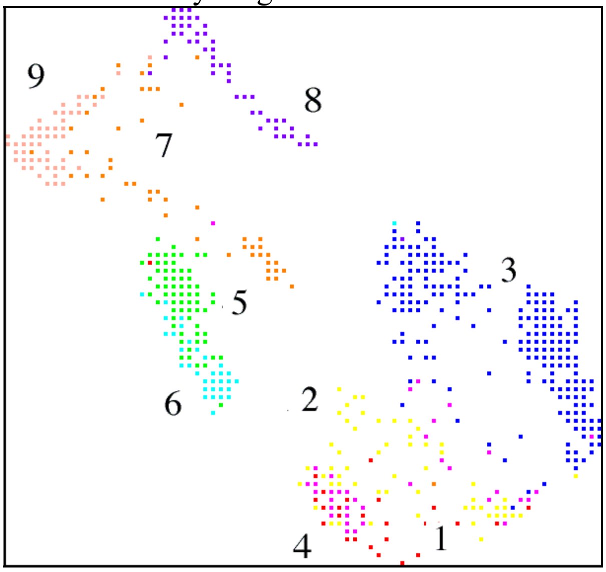

Figure 10 Distribution of DataBots on Olive Oil Dataset after 400 steps.

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. I.e. it can be predicted that the consistency of the olive oils vary according to the producing regions.

9 Conclusion

In this work we describe a novel approach to Data Mining using autonomous minirobots. To our knowledge this is one of the first attempts to use autonomous minirobots 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 [9]. It could be shown that a very simple anatomy and very simple strategies lead to surprising results concerning the detection of clusters and preserving the overall structure of high dimensional datasets.

This approach is in several senses a very natural one. First of all, it makes use of a metaphor that is attractive to both, researchers and clients in order to learn from nature the ability to detect structures in novel worlds. In the field of Data Mining these worlds consist of high dimensional datasets. Secondly, for emergence it is absolutely necessary that a huge number of elementary processes cooperate. A new level or niveau can only be observed when elementary processes are disregarded and only the overall structures, i. e. structures formed by the cooperation of many elementary processes, are regarded. This calls for implementation on massively parallel systems. 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 simualtion tool that uses autonomous minirobots for the evaluation of high dimensional data structures [9].

in: Polani, D. et al.(eds): Abstracting and Synthesizing the Principles of living Systems, GWAL, Lübeck pp 191-199

Acknowledgements

The author wishes to thank J. Gasteiger from University of Erlangen-Nuremberg and Prof. Fiorna 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 now current third version is currently developed by Dirk Malorny

References

[1] Haken, H.: Synergetics, an Introduction, Springer, Berlin 1974

[2] Kohonen, T.: Self-Organized Formation of Topologically Correct FeatureMaps, Biological Cybernetics Vol. 43, pp 59 - 69, 1982

[3] Malorny, D., Müller,I., Münchberg, F.: vv, Technical Note, Department of Computer Science, University of Marburg, Hans-Meerwein- Str., 35032 Marburg, 25. Apr. 1998

[4] Ultsch, A.: Self-organizing Neural Networks for Visualization and Classification, in O. Opitz, B. Lausen and R. Klar, (Eds.) Information and Classifi-cation, Berlin: Springer-Verlag, 307-313.

[5] Ultsch, A:: The Integration of Neural Networks with SymbolicKnowledge Processing, in Diday et al. „New Approaches in Classification and Data Analysis“, pp 445 - 454, Springer Verlag 1994

[6] Ultsch, A.: Self-Organizing Neural Networks Perform Different from Statistical k-means clustering, Gesellschaft f. Klassifikation, Basel 8th - 10th March, 1995

[7] Ultsch, A.: 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

[8] Ultsch, A: Umgebungs-Dynamik-Universen, Technical Note, Department of Computer Science, University of Marburg, Hans-Meerwein- Str., 35032 Marburg, 25. Apr. 1998.

[9] Ultsch, A: Data Mining and Knowledge Discovery with Self-Organzing Feature Maps for Multivariate Time Series,in Oja,E., Kaski,S.: Kohonen Maps, p 33- 46, Elsevier, 1999.

[10] 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