Data Mining and Knowledge Discovery with Emergent Self-Organizing Feature Maps for Multivariate Time Series

A. Ultsch
Philipps-University of Marburg
Department of Computer Science
Hans-Meerwein-Str
35032 Marburg
Germany

Self-Organizing Feature Maps, when used appropriately, can exhibit emergent phenomena. SOFM with only few neurons limit this ability, therefore Emergent Feature Maps need to have thousands of neurons. The structures of Emergent Feature Maps can be visualized using U-Matrix Methods. U-Matrices lead to the construction of self-organzing classifiers possessing the ability to classify new datapoints. This subsymbolic knowledge can be converted to a symbolic form which is understandable for humans. All these steps were combined into a system for Neuronal Data Mining. This system has been applied successfully for Knowledge Discovery in multivariate time series.

1 Introduction

Data Mining aims to discover so far *unknown* knowledge in large datasets. The most important step thereby is the transition from subsymbolic to symbolic knowledge. Self-Organizing Feature Maps are very helpful in this task. If appropriately used, they exhibit the ability of emergence. I. e. using the cooperation of many neurons, Emergent Feature Maps are able to build structures on a new, higher level. The U-Matrix-method visualizes these structures corresponding to structures of the high-dimensional input-space that otherwise would be invisible. A knowledge conversion algorithm transforms the recognized structures into a symbolic description of the relevant properties of the dataset.
In chapter two we shortly introduce our approach to Data Mining and Knowledge Discovery, chapter three clarifies the use of Self-Organizing Feature Maps for Data Mining. Chapter four clarifies the use of Feature Maps in order to obtain emergence. In the chapters five and six those steps of Data Mining, where Feature Maps can be used, are described. Chapter seven is a description of our system, the so called Neuronal Data Mine which uses Emergent Feature Maps and Knowledge Conversion. In chapter eight an important application area - Knowledge Discovery in multivariate time series - is described. Chapter nine gives first results of an application of this system.

2 Data Mining and Knowledge Discovery

Since the use of the term Data Mining is quite diverse we give here a short definition in order to specify our approach to Data Mining and Knowledge Discovery. A more detailed description can be found in [Ultsch 99a]. We define Data Mining as the inspection of a large dataset with the aim of Knowledge Discovery. Knowledge Discovery is the discovery of new knowledge, i. e. knowledge that is unknown in this form so far. This knowledge has to be represented symbolically and should be understandable for human beings as well as it should be useful in knowledge-based systems. Central issue of Data Mining is the transition from data to knowledge. Symbolically represented knowledge - as sought by Data Mining - is a representation of facts in a formal language such that an interpreter with competence to process symbols can utilize this knowledge [Ultsch 87]. In particular, human beings must be able to read, understand and evaluate this knowledge. The knowledge should also be useable by knowledge-based systems.

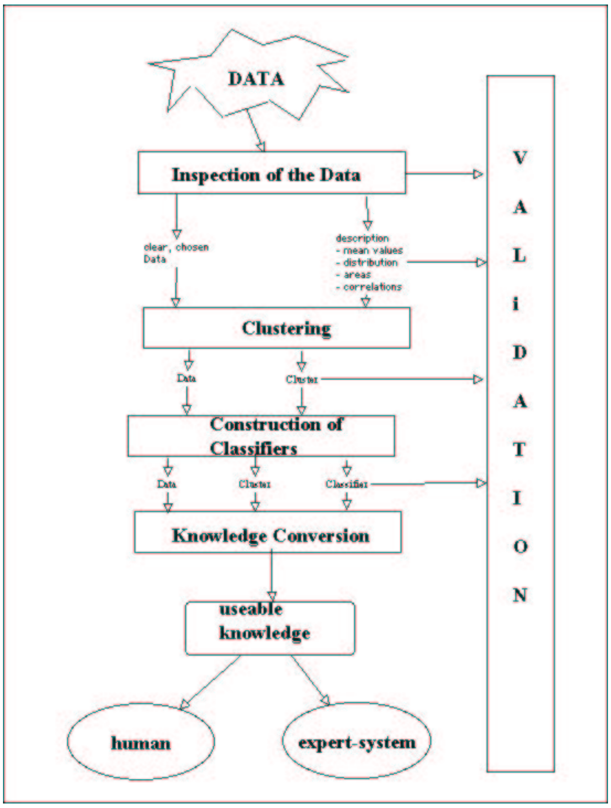

Figure 1: Steps of Data Mining

The knowledge should be useful for analysis, diagnosis, simulation and/or prognosis of the process which generated the dataset. We call the transition from data, respectively an unfit knowledge representation, to useful symbolic knowledge Knowledge Conversion [Ultsch 98].
Data Mining can be done in the following steps:

- inspection of the dataset
- clustering
- construction of classifiers
- knowledge conversion and
- validation (see figure 1 for an overview)

Unfortunately it has to be stated that in many commercial Data Mining tools there is no Knowledge Conversion [Gaul 98]. The terms Data Mining and Knowledge Discovery are often used in those systems in an inflationary way for statistical tools enhanced with a fancy visualization interface [Woods, Kyral 98].
The difference between exploratory statistical analysis and Data Mining lies in the aim which is sought. Data Mining aims at Knowledge Discovery.

3 SOFM for Data Mining

Figure 1 shows the different steps in Data Mining in order to discover knowledge. Statistical techniques are commonly used for the inspection of the data and also for their validation. Self-Organizing Feature Maps can be used for classification and the construction of classifiers. Particularly well suited for these tasks are Emergent Feature Maps as described in the next chapter. Classifiers constructed with the use of a Self-Organizing Feature Map do, however, not possess a symbolic representation of knowledge.

They can be said to contain subsymbolic knowledge. In the step Knowledge Conversion the extraction of knowledge from Self-Organizing Feature Maps will be performed. One method to extract knowledge from Self-Organizing Feature Maps is the so called sig*-algorithm which will be briefly described in chapter six.

4 Emergent vs. Non-emergent Feature Maps

Self-Organizing Feature Maps were developed by Teuvo Kohonen in 1982 [Kohonen 82] and should, to our understanding, exhibit the following interesting and non-trivial property: the ability of emergence through self-organization. Self-organization means the abilty of a biological or technical system to adapt its internal structure to structures sensed in the input of the system. This adaptation should be performed in such a way that first, no intervention from the environment is necessary (unsupervised learning) and second, the internal structure of the self-organizing system represents features of the input-data that are relevant to the system. A biological example for self-organization is the learning of languages by children. This process can be done by every child at a very early age for different languages and in quite different cultures.

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 “phase-change”. 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, etc.

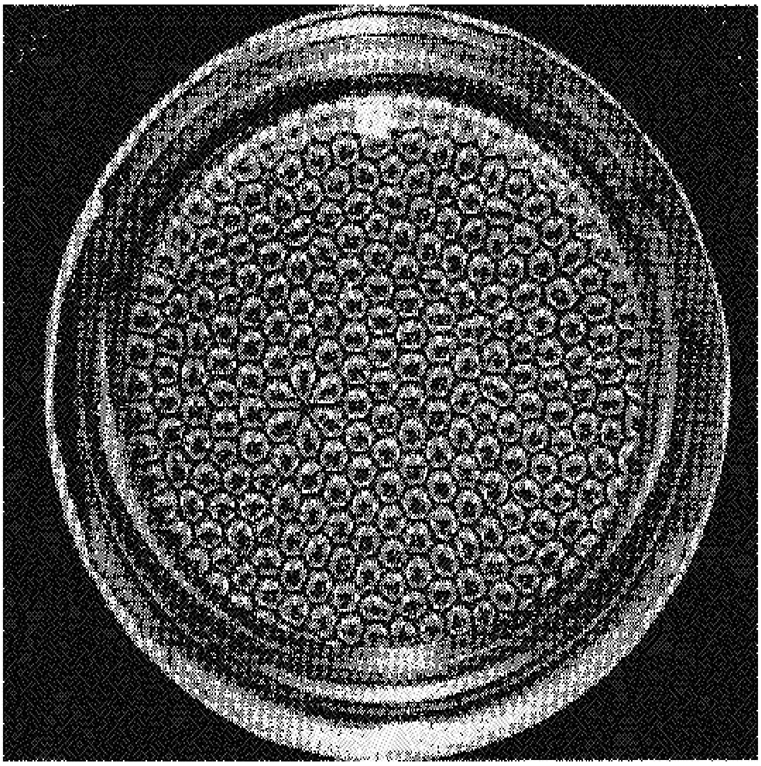

Fig. 2: Hexagonal convection cells on a uniformly heated copper plate [Haken 71]

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 form 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 (elemantary processes)

produce a coherent radiation beam.
Although Kohonen's Self-Organizing Feature Maps are able to exhibit emergence they are often used such, that emergence is impossible. For emergence it is absolutly 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 considered.
In typical applications of Kohonen's Self-Organizing Feature Maps the number of neurons in the feature maps are too few to show emergence. A typical example, which is representative for many others, is taken from [Reutterer 99]. The dataset describes consumers of household goods. Each household is described by a nine-dimensional vector of real numbers. Self-Organizing Feature Maps are used to gain some inside-knowledge in the structure and segmentation of the market for the household goods. A Self-Organizing Feature Map with three by three, i. e. nine neurons has been used [Reutterer 99]. Using Kohonen's learning scheme, each of the nine neurons represents the bestmatch of several input-data. Each neuron is considered to be a cluster.
Common to all non-emergent ways to use Kohonen's feature maps is that the number of neurons is roughly equal to the number of clusters expected to be found in the dataset. A single neuron is typicaly regarded as a cluster, i. e. all data, whose bestmatches fall on this neuron, are members of this cluster. It seemed for some time that this type of Kohonen's feature maps performs clustering in a way that is similar to a statisical clustering algorithm called k-means[Ultsch 95].
An absolutely necessary condition for emergence is the cooperation of many elementary processes. Emergence is therefore only expected to happen in Self-Organizing Feature Maps with a large number of neurons. Such feature maps, we call them Emergent Feature Maps, have typically at least some thousands if not tens of thousands of neurons. In particular the number of neurons may be much bigger than the number of datapoints in the input-data. Consequently most of the neurons of Emergent Feature Maps will represent very few input-points if any. Clusters are detected on Emergent Feature Maps not by regarding single neurons but by regarding the overall structure of the whole feature map. This can be done by using U-Matrix-methods [Ultsch 94]. With Emergent Feature Maps we could show that Self-Organizing Feature Maps are different and often superior to classical clustering algorithms [Ultsch 95]. A canonical example, where this can be seen, is a dataset consisting of two different subsets. These subsets are taken from two well seperated toroids that are interlinked like a chain as it can be seen in figure 3.
Using an Emergent Feature Map of a dimension 64 by 64 = 4096 neurons, the two seperate links of the chain could easily be distinguished. In contrast to this, many statistical algorithms, in particular the k-means algorithm, were unable to produce a correct classification. We think that the task of Data Mining, i. e. the seeking of new knowledge, calls for Emergent Feature Maps. The property of emergence, i. e. the appearance of new structures on a different abstraction level, coincides well with the idea of discovering new knowledge.

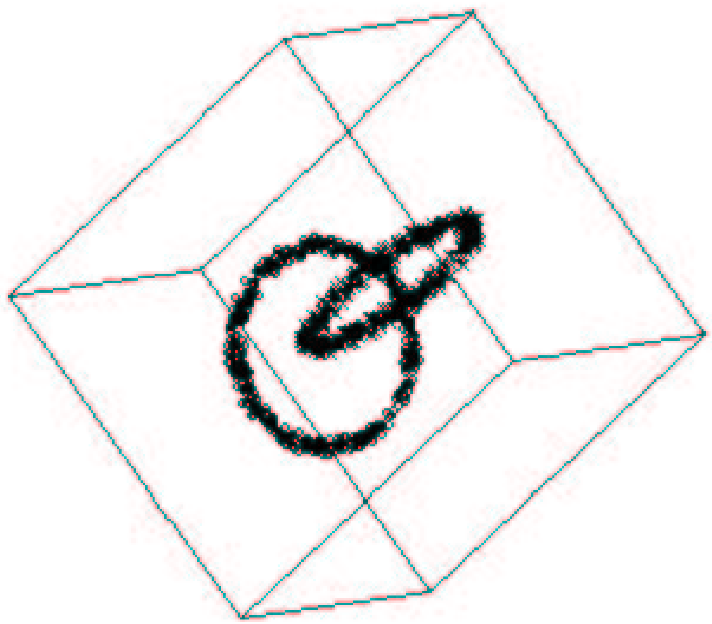

Fig. 3: Chainlink Dataset

5 Construction of Classifiers

When Emergent Feature Maps with a sufficiently large number of neurons are trained with high-dimensional input-data, these datapoints distribute sparsely on the feature map. Regarding the position of the bestmatches, i. e. those neurons, whose weights are most similar to a given input-point, gives no hint on any structure in the input-dataset. In the following picture a three-dimensional dataset consisting of thousand points was projected on a 64 by 64 Emergent Feature Map. The topology of the feature map is toroid, i. e. the borders of the map are cyclically connected. The positions of the bestmatches exhibit no structure in the input-data.
In order that structures of the input-data can emerge we use the so called U-Matrix-Method. The simplest of these methods is to sum up the distances between the neurons-weights and those of its immediate neighbours. This sum of distances to its neighbours is displayed as elevation at the position of each neuron. The elevation-values of the neurons produce a three-dimensional landscape, the so called U-Matrix. U-Matrices have the following properties:

- Bestmatches that are neighbours in the high-dimensional input-data space lie in a common valley.
- If there are gaps in the distibution of input-points, hills can be seen on the U-Matrix. The elevation of the hills is proportional to the gap distance in the input-space.
- The principal properties of Self-Organizing Feature Maps, i. e. conserving the overall topology of the input-space, is inherited by the U-Matrix. Neighbouring data in the input-space can also be found at neighbouring places on the U-Matrix.
- Topological relations between the clusters are also represented on the two-dimensional layout of the neurons

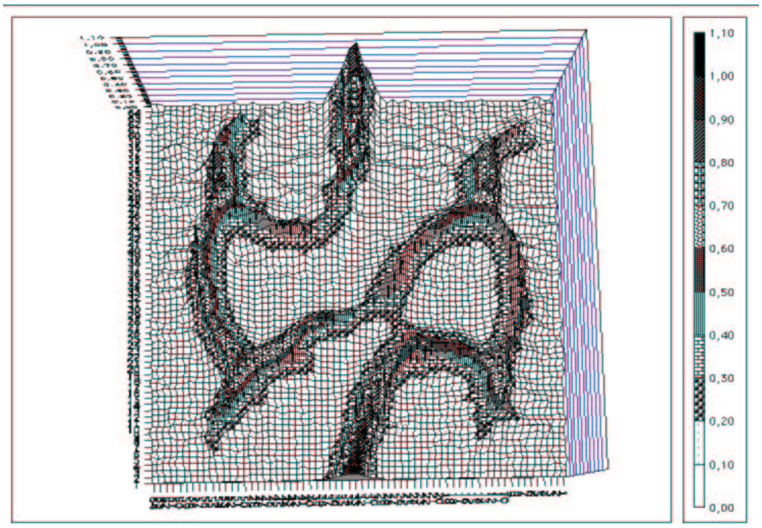

Fig. 4: An U-Matrix of the Data

With U-Matrix-Methods emergence in Kohonen maps has been observed for many different applications, for example: medical diagnosis, economics, environmental science, industrial process control, meteorology, etc.

The cluster-structure of the input-dataset is detected using an U-Matrix. Clusters in the input-data can be detected in the U-Matrix as valleys surrounded by hills with more or less elevation, i. e. clusters can be detected, for example, by raising a virtual waterlevel up to a point, where the water floods a valley on the U-Matrix. Regarding an U-Matrix the user can indeed grasp the high-dimensional structure of the data. Neurons that lie in a common valley are subsumed to a cluster. Regions of a feature map that have high elevations in an U-Matrix are not identified with a cluster. Neurons that lie in a valley but are not bestmatches are interpolations of the input-data. This allows to cluster data with Emergent Feature Maps. This approach has been extensively tested over the last years and for many different applications. It can be shown that this method gives a very good picture of the high-dimensional and otherwise invisible structure of the data. In many applications meanings for clusters could be detected.

Emergent Feature Maps can be easiliy used to construct classifiers. If the U-Matrix has been seperated into valleys corresponding to clusters and hills corresponding to gaps in the data, then an input-datapoint can be easily classified by looking at the bestmatch of this datapoint. If the point's bestmatch lies inside a cluster-region on the U-Matrix the input-data is added to that cluster. If the bestmatch lies on a hill in the U-Matrix, no classification of this point can be assigned. This is in particular the case if the dataset possesses new features, i. e. aspects that were not included in the data learned so far. With this approach, for example, outliers and errouneous data are easily detected.

6 Knowledge Conversion

The classifiers constructed with Emergent Feature Maps and the U-Matrix described in the last chapter, possess the "knowledge" to classify new data. This knowledge is, however, not symbolic. Neither a reason, why a particular dataset belongs to a particular cluster, nor, why a given dataset can not be classified, can be given. What is necessary at this point is to convert this type of knowledge to a symbolic form.

We have developed an algorithm called sig* in order to perform this Knowledge Conversion [Ultsch 94]. As input sig* takes the classifier as described in the last chapter. A symbolic description of all the weights of the neurons belonging to a particular cluster is constructed. Sig* generates description using decision-rules. These decision-rules contain as premises conditions on the input-data and as conclusions the decision for a particular cluster. Clusters are described by two different types of rules. There are so called characterization rules which describe the main characteristics of a cluster. Secondly, there are rules which describe the difference between a particular cluster and neighbouring clusters. The different steps of this Knowledge Conversion can be described as follows:

- Selection of components of the high-dimensional input-data that are most significant for the characterization of a cluster
- Construction of appropriate conditions for the main properties of a cluster
- The Composition of the conditions in order to produce a high-level significant description.

To realize the first step, sig* uses a measure of significance for each component of the high-dimensional input-data with respect to a cluster. The algorithm uses only very few conditions, if the clusters can be easily described. If the clusters are more difficult to describe sig* uses more conditions and more rules to describe the differences are generated in order to specify the borders of a cluster. For the representation of conditions, sig* uses interval-descriptions for characterization-rules and splitting-conditions for the differention-rules. The conditions can be combined using "and", "or" or a majority-vote. It could be shown that for known classifications sig* reproduces 80 to 90 ++ % of the classification-ability of an Emergent Feature Map.

7 The Neuronal Data Mine NDM

The methods presented in the previous chapters have been developed and refined over the last years and combined to a tool for Data Mining and Knowledge Discovery called Neuronal Data Mine [Ultsch 99a]. This tool contains the following modules :

- Statistics (Inspection of the Data)
- Emergent Feature Maps (Clustering)
- U-Matrix (Construction of Classifiers)
- sig* (Knowledge Conversion)
- Validation

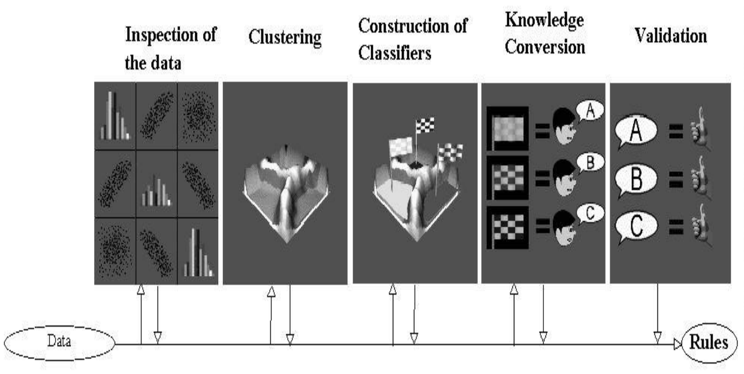

Fig. 5: Screen-shot of the user-interface of the NDM

8 The Neuro Data Mine for Knowledge Discovery in Time Series

One of the latest and most fascinating applications of NDM is Data Mining in multivariate time series. The key for this application is a suitable knowledge representation for temporal structures (see chapter 2). With the definition of unification-based temporal grammars (UTG) this key-issue has been solved [Ultsch

99b]. UTGs belong to the class of definitve clause grammars. UTGs describe temporal phenomena by a hierachy of semiotic description-levels. Each semiotic description-level consists of a symbol (syntax), a description (semantic) and an explanation useful in the context of a particular application (pragmatic). Temporal phenomena are described on each level using temporal phases and temporal operations. The latter are called connexes. As phases we identified Primitive Patterns, Successions, Events, Sequences and Temporal Patterns. Primitive Pattern represent the different elementary conditions of the process described by the multivariate time series. Successions model the duration, Events the simultanety of temporal phases. Sequences are used to formulate repetitions. Temporal Patterns finally condense variations in Sequences to a common abstract description of important temporal patterns. The phases described above can be combined using connexes for duration, simultanety and temporal sequence. These temporal operations are designed to be flexible with regard to time. The connexes require not a coincidence of events in a mathematically sense. They allow a certain flexibility, i. e. events that are sufficiently close in time are considered to be simultaneous. This is necessary since the multivariate time series stem from natural or technical processes that have always a certain variation in time. A special fuzzy representation was used to represent this flexibility [Ultsch 99b]. This approach leads to only three temporal operations for the representation of temporal features. In other representation formalisms, for example Allen, much more temporal operations are necessary [Allen 84].
Emergent feature maps are used for Temporal Data Mining in the following steps:

- description of the elementary conditions of the process (Primitive Pattern)
- description of the duration of phases (Successions)
- description of simultaneity (Events)
- detection of start- and end-points of temporal patterns (Sequences resp. Temporal Patterns).

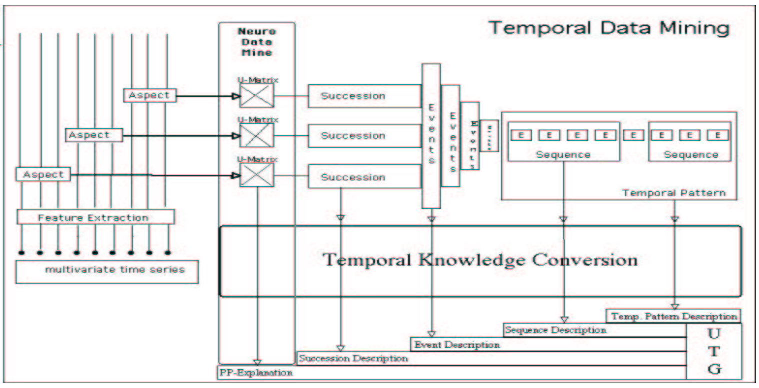

Fig. 6: Temporal Data Mining

9 Application: Sleep Related Breathing Disorders

In a first application the Temporal Data Mine has been used for a medicalproblem, the so called Sleep Related Breathing Disorders (SBRD) [Penzel et al 91]. Humans who suffer from SRBD experience the stopping of breathing during certain periods of sleep. The stopping-periods are critical if they last at least 10 seconds and occur more than 40 times per hour [Penzel et al 91]. As multivariate time series were considered: EEG, EMG, EOG, EKG, airflow, thorax- and abdomen movements, airflow and saturation of blood with oxygen.

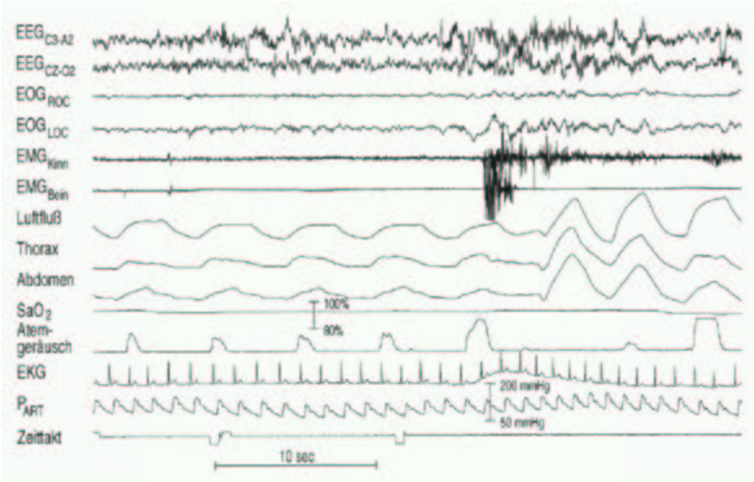

Fig. 7: Multivariate Time Series for SRBD

For those multivariate time series, two different types of U-Matrix called "air" and "move", were generated [Guimaraes/Ultsch 99]. The U-Matrix "air" focuses on all aspects of the time series related to airflow. The U-Matrix "move" concentrates on aspects of movements of thorax and abdomen. In the "air"-U-Matrix six elementary states were identified. These elementary states (clusters) are considered elementary primitive temporal elements and termed Primitive Patterns. In the U-Matrix "move" nine Primitive Patterns could be identified. The temporal sequence of the Primitive Pattern are represented as paths on the U-Matrices. Using temporal knowledge conversion, six Events and five different Temporal Patterns could be found in the time series. The knowledge was formulated in UTG notation. All semiotic description levels of the UTG (see last chapter) have been presented to an expert in order to evaluate the plausibility of the phases and the descriptions. This showed that all the events found represented important medical properties. In particular the events could be related to physiological stages like, for example, "obstructive snoring" or "hyperpnoe". Four of the five Temporal Patterns that were discovered were very well known to the expert and could be assigned a medical meaning. One of the Temporal Patterns was a newly discovered pattern inherent in some type of human sleep. This gave a hint on a potential new way to look onto this certain types of sleeping disorders. In the following picture an example of a Temporal Pattern and the corresponding multivariate time series is depicted.

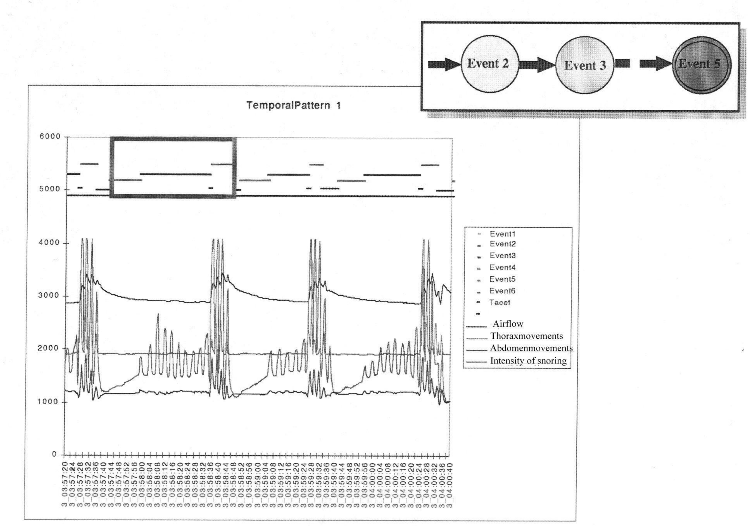

Fig 8: Temporal Knowledge

10 Conclusion

In Data Mining the first step after the inspection of a dataset is the identification of clusters. SOFM with only few neurons limit implicitly the number of clusters to be found in the data. With such feature maps only a very crude insight into the input-data can be gained, if at all. In feature maps possessing thousands of neurons, U-Matix-methods can be used to detect emergence. U-Matrices visualize structures in the data by considering the cooperation of many neurons. The structures seen give insights to the otherwise invisible high-dimensional dataspace. It can be shown that emergent feature maps are superior to other clustering methods, particularly to k-means [Ultsch 95].
The most important step of Data Mining is Knowledge Conversion, i. e. the transition from a subsymbolic to a symbolic representation of knowledge. Emergent feature maps provide an excellent starting-point for Knowledge Conversion. Other classifiers such as decision-trees focus on the efficiency of the discrimination between clusters. Declarative rules, extracted from U-Matrices, using sig* provide an extract description of significant properties of clusters.
The methods described above could be used to analyze multivariate time series. Unification-based grammars (UTG) have been developed as a tool to represent symbolic knowledge for Temporal Data Mining. The approach has been successfully tested for a medical problem regarding sleep disorders.

11 References

[Allen 84] Allen, J.: *Towards a General Theory of Action and Time, Artifical Intelligence* 23, 1984, S 123 - 154

[Gaul 98] Gaul, W. G. *Classification and Positioning of Data Mining Tools,* Herausforderungen der Informationsgesellschaft an Datenanalyse und Wissensverarbeitung 22. Jahrestagung Gesellschaft für Klassifikation

[Guimarães/Ultsch 99] Guimarãess, G. Ultsch, A.: *A Method for Temporal Knowledge Conversion*, to appear.

[Guimarães/Ultsch 96] Guimarães, G. Ultsch, A.: *A Symbolic Representation for Pattern in Time Series Using Definitive Clause Grammars*, 20 th Annual Conference of the Society for Classification, Freiburg 6th - 8th March 1996, pp 105 - 111

[Kohonen 82] Kohonen, T.: *Self-Organized Formation of Topologically Correct Feature Maps,* Biological Cybernetics Vol. 43, pp 59 - 69, 1982

[Penzel et al. 91] Penzel, P., Stephan, K., Kubicki, S., Herrmann, W. M.:*Integrated Sleep Analysis with Emphasis on Automatic Methods,* In: R. Degen, E. A. Rodin (Eds.) Epilepsy, Sleep and Sleep Deprivation, 2nd ed. (Epilepsy Res. Suppl. 2), Elsevier Science Publisher, 1991, S 177 - 2041

[Reutterer 98] Reutterer; T, Panel *Data Based Competitive Market Structure and Segmentation Analysis using Self-Organizing Feature Maps,* Proc.Annual Conf. Siciety for Classification, pp 92, Dresden, 1998

[Ultsch 99a] Ultsch, A.:*Data Mining und Knowledge Discovery mit Neuronalen Netzen,* Technical Report, Department of Computer Science, University of Marburg, Hans-Meerwein-Str., 35032 Marburg

[Ultsch 99b] Ultsch, A.:*Unifikationsbasierte Temporale Grammatiken für Data Mining und Knowledge Discovery in multivariaten Zeitreihen,* Technical Report Department of Computer Science, University of Marburg, March 1999

[Ultsch 98] 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

[Ultsch 95] Ultsch, A.: *Self-Organizing Neural Networks Perform Different from Statistical k-means clustering,* Gesellschaft für Klassifikation, Basel 8th - 10th March, 1995

[Ultsch 94] Ultsch, A.: *The Integration of Neural Networks with Symbolic Knowledge Processing,* in Diday et al. "New Approaches in Classification and Data Analysis", pp 445 - 454, Springer Verlag 1994

[Ultsch 87] Ultsch, A.: *Control for Knowledge-based Information Retrieval,* Verlag der Fachvereine, Zürich, 1987

[Woods/Kyral 97] Woods, E., Kyral, E. *Data Mining,* Ovum Evaluates, Catalumya Spain, 1997