

Unsupervised System for Discovering Patterns in Time-Series

Alexandra Scherbart¹ and Nils Goerke²

¹ Applied Neuroinformatics Group, Faculty of Technology,
Bielefeld University, Postfach 100131, 33501 Bielefeld, Germany,

`ascherba@techfak.uni-bielefeld.de`,

² Div. of Neural Computation, Dept. of Computer Science,

University of Bonn, Germany,

`{scherbart,goerke}@nero.uni-bonn.de`

ABSTRACT: Within this paper we present a framework for discovering patterns in time-series by unsupervised feature selection and unsupervised, self-organised clustering. The proposed unsupervised feature selection algorithm is determining the feature relevance for a variety of transformations to select a set of features to build the feature space. We propose to take the phase space as basis and extend it by the selected features. The core of the presented approach is the self-organised clustering in feature space with Multi-SOMs and Multi-Neural-Gas. For lack of prior knowledge about the real number of clusters, K is estimated by using a combination of common clustering analysis coefficients that have been adapted for M-SOMs and M-NGas. The presented approach of unsupervised feature selection, feature space construction and subsequent unsupervised clustering has been evaluated with time series from a robotic application with promising results.

KEYWORDS: Unsupervised, Self-Organised Clustering, Multi-SOMs, Multi-Neural Gas, Clustering Analysis

1 INTRODUCTION

There is an urgent need in extracting useful information within data and making it understandable for humans. Data Mining, a step in the Knowledge Discovery process, is trying to meet these challenges. Data Mining in time-series includes the steps of data analysis and clustering.

Major tasks in Pattern Discovery and Pattern Recognition in time-series are about finding a “good” feature space. Feature Selection is the task of getting smaller sets of representative features. Existing approaches for Feature Selection involve maximum entropy based methods, expectation maximization (EM), etc. [11].

Having selected a set of features, the next step is to determine a grouping within given unlabeled data, that is representing the underlying data distribution well. Unsupervised or self organising clustering apply in the case that no preclassified data points are available. Classifying data points and finding clusters within data sets is an application where neural networks have been successfully applied.

Within the work presented, the Multi-Self-Organising-Map (Multi-SOM) [1] and Multi-Neural-Gas approach [3, 4] have been applied for self organised clustering of unlabeled data.

2 APPROACH

In this work we are presenting an unsupervised system to discover patterns in time-series including Feature Selection and Self-Organised Clustering. Under the presumption that no prior knowledge about the time-series and application is given, there is no knowledge about inferable suitable features and existing patterns.

Instead of defining or searching for one fixed feature space, we followed up the idea to let the system select which representative features to take. A desirable, suitable feature space would be one, in which the characteristics of the underlying time-series can be separated well. Features carrying little or no additional information are redundant and should be discarded. We propose to take the phase space as a basis for the feature space and carefully select additional features (Feature Selection) to build the feature space for subsequent clustering. Motivated by the theory of nonlinear dynamical systems [13] the phase space has been used because it can reflect the true dynamics of the system that have generated the time-series. From the variety of implemented

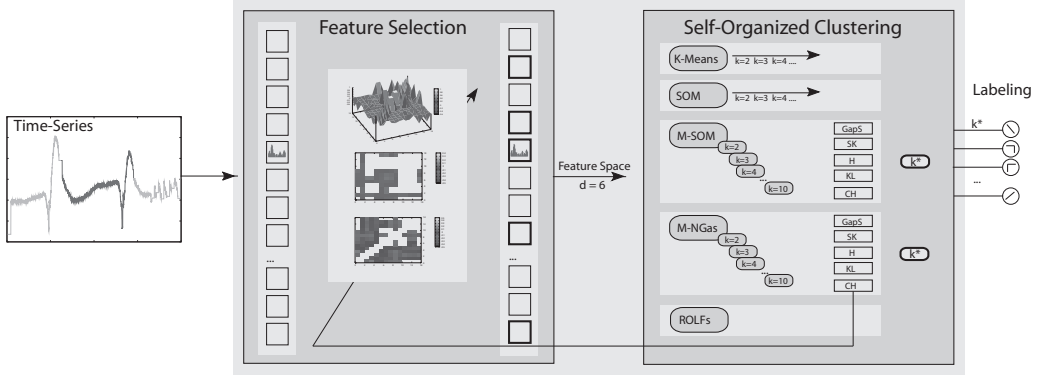


Figure 1: Sketch of the System

transformations the proposed heuristic driven feature selection mechanism selects those features (f_i) that show the maximal feature relevance (r_i).

3 IMPLEMENTATION

3.1 TRANSFORMATIONS AND INDUCED FEATURE SPACES

A set of transformations has been chosen and implemented to map the time-series into the feature space induced by this very transformations. From the variety of imaginable transformations a total of 19 have been chosen, which belong to three different categories A) frequency domain based transformations, B) time domain based transformations, C) phase space, reconstructed phase space.

The following transformations have been implemented because the authors believe that these transformations are capable to extract valuable features from the time-series: from the frequency domain: short-time-fourier-transformation (STFT), from the theory of multi media retrieval and audio processing *Timbral Texture Features* proposed by [15]: *Spectral Centroid*(SC), *Spectral Flux*(SF), *Spectral Rolloff*(SR), *Mel Frequency Cepstrals*(MFC), *RMS*(RMS) and *Time Domain Zero Crossings*(TDZC). Further realised features are based on windowed transformations. The window size was set to $w = 256$ and shifted by an offset of 32. We provided *Integral*(Int), *number of inflection points*(IP), the first four statistic moments *Expectation*(Mean), *Variance*(Var), *Skewness*(Sk) and *Kurtosis*(Ku), the evaluated Autocorrelation-function features *First Zero Crossing*(FZC) and *Number of Zero Crossings*(NZC).

Based on results from the theory of nonlinear dynamical system the *phase space*, and the *reconstructed phase space* are promising feature space components. The 3-dimensional phase space is built by the time-series $x(t)$ and their first and second derivative ($x(t), \dot{x}(t), \ddot{x}(t)$). The reconstructed phase space is built by the delayed variables ($x(t), x(t + \tau), x(t + 2 \cdot \tau)$), where τ is the first zero crossing of the autocorrelation function.

3.2 FEATURE SELECTION

From the variety of chosen features f_i , only a subset of features is used to build the reduced feature space \mathcal{F}^* to concentrate on those features that are most valuable. As an additional benefit, the dimensionality of the feature space is thereby further reduced for getting an easier treatment for the clustering procedure.

We propose to select those features, that are the most unlike for carrying the maximum of information into the reduced feature space \mathcal{F}^* . Therefore we have calculated the covariance $\text{cov}(f_i, f_j)$ and the correlation $\text{corr}(f_i, f_j)$ between all features f_i available, with respect to the given time-series or time-series fragments. The correlation, and covariance matrices are depicted in Fig. 2. Values within these correlation and covariance matrices that are close to zero give a good hint that these features are independent of each other.

To determine which features f_i are the most relevant, we define the feature relevance r_i as the quantity of correlation and covariance values that are close to zero, that is below a threshold of $b = 0.1$. Thus, these features f_i maximise the relevance build the reduced feature space \mathcal{F}^* .

$$r_i = \sum_{j=1}^{j=N} \Theta(\text{corr}(f_i, f_j)) + \Theta(\text{cov}(f_i, f_j)) \quad \text{with} \quad \Theta(c) = \begin{cases} 1 & : |c| \leq b \\ 0 & : \text{else} \end{cases} \quad (1)$$

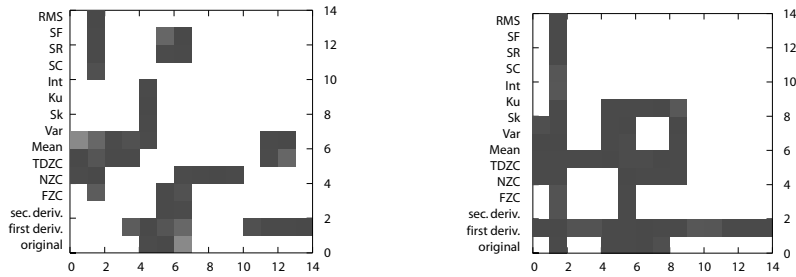


Figure 2: Correlation (left) and Covariance (right) matrices for evaluated features of time-series I (see Sec. 4.1). Only the entries of matrices with relevance values $|\text{corr}(\cdot)| \leq 0.1$ and $|\text{cov}(\cdot)| \leq 0.1$ are shown.

3.3 SELF-ORGANISED CLUSTERING AND CLASSIFICATION

Multi-SOMs

Multi-SOMs are an extension of the Self-Organising Map [7] approach to a set of several neural networks. Each of these K partner SOMs is treated as an individual of its own. The learning procedure is slightly deviated: only the winner SOM is adapted following the normal SOM learning rule. The winner SOM is that partner-SOM to which the winning neuron belongs to. Except for that winner SOM all other, non-winning SOMs are kept unchanged.

As for classical SOMs, the learning rate η and the size of the neighborhood $h(t, i, n)$ should decrease in time [8, 9]. Depending on the number of epochs t_{max} the learning rate η is decaying from initial value η_i to a final value η_f and the size of the neighborhood σ decays from σ_i to σ_f :

$$\eta(t) = \eta_i \cdot \left(\frac{\eta_f}{\eta_i} \right)^{\frac{t}{t_{max}}}, \quad \sigma(t) = \sigma_i \cdot \left(\frac{\sigma_f}{\sigma_i} \right)^{\frac{t}{t_{max}}}$$

During the learning process, the mean deviation between the pattern \mathbf{P} and the winning neuron \mathbf{V}_i is used to monitor the progress of learning. Good results were found for rapid learning with $\eta_i = 0.1$, $\eta_f = 0.01$, $\sigma_i = 2.0$ and $\sigma_f = 0.1$ for $t \leq t_{max}$ and keeping $\eta = \eta_f = 0.01$, $\sigma = \sigma_f = 0.1$ fixed for the rest of the learning process ($t_{max} \leq t \leq 2 \cdot t_{max}$).

Multi-Neural-Gas

In the extended approach of the Neural-Gas [9] to a Multi-Neural-Gas several disconnected neural networks are treated individually. As for Multi-SOMs the learning rule is altered in such a way, that only the winner neural gas network, where the winning neuron belongs to, is adapted. Following the proposal of [9] the parameters $\eta_i = 0.5$, $\eta_f = 0.01$, $\lambda_i = 10$ and $\lambda_f = 0.5$ for each Neural-Gas network were used to reach a good adaption and convergence of the neurons to the data. Similar to Multi-SOMs, η is decaying from η_i to η_f (same for λ) for $t \leq t_{max}$ and kept constant $\eta = \eta_f = 0.01$, $\lambda = \lambda_f = 0.1$, if $t_{max} \leq t \leq 2 \cdot t_{max}$. After convergence, each of these individual partner networks is regarded as a cluster of its own.

Thus, Multi-SOMs and Multi-Neural Gas are not only adapting to the underlying data distribution, but are directly yielding a set of K clusters.

Estimating the number of clusters

For lack of any prior knowledge of the underlying data distribution an optimal number K^* of clusters and therefore the number of partner-networks has to be estimated. But, as for other clustering techniques, the number of partner networks has to be determined in advance. To overcome this, over varied number of partner-networks K is iterated. Having clustered with e. g. $K = 1, \dots, 10$ partner-SOMs and partner-NGas respectively, an estimation is done by the quality measures *Silhouette Coefficient* and *Gap Statistics* to yield an optimal number K^* . These quality measures have been extended for M-SOMs and M-NGas [3, 4]. The decision for the best estimated number of clusters is done by rule of majority vote. The statistics yielding the most definite decision is taken for determining K^* .

The next step to consider would be to make this part, i. e. the estimation based on the results of clustering-quality measures, learnable, e. g. by Reinforcement Learning.

Classification

The classification procedure is divided into 4 subtasks.

- A *Find a set of classes that represents the given set of data points in an adequate way.* This is performed by adaptation of the networks to the data distribution directly by the M-SOM and M-NGas learning rule.
- B *Assign a given data point to that class, that is representing this data point best.* This is performed by the neurons, directly by finding the winning neuron i , and thereby finding the winning-network.
- C *Find a typical representative of a given class.* The task is to find typical patterns in input space (time-series) that are typical for the classes determined. Since an inversion of the performed calculations, from feature space to input space is in general impossible, or at least very complicated, we propose a different way. Find those neurons that represent their clusters best, and seek for input patterns that maximise the response of those very neurons. For the M-SOMs the neurons in the center of each partner-SOM can be taken as representative neurons. A different strategy applies for M-NGas and M-SOMs as well: take that neuron, that is closest to the most dense part of the data distribution.
- D *Labeling* The classes have to be labeled according to the found patterns which means, to assign human understandable labels or symbols to the found classes with respect to a given application.

4 RESULTS OF EXPERIMENTS

We have conducted several series of experiments with a realistic but artificial and real (robot-) sensory time-series. Since the Multi-Neural-Gas showed to be superior, all presented results were performed with a Multi-NGas of K-25 type (25 neurons per NGas). The parameters for the learning process are adjusted as described above (sec. 3.3) to $\eta_i = 0.5$, $\eta_f = 0.01$, $\lambda_i = 10$, $\lambda_f = 0.5$ and $t_{max} = 20$, i.e. learning is done over 40 epochs at all. With respect to different numbers $K = 1, \dots, 10$ of partner Neural-Gas the quality measures Mean Deviation, Silhouette coefficient and Gap Statistics, have been evaluated for each iteration.

To overcome any dependencies from an unfavourable random starting initialisation of the neurons, we conducted several runs with different initialisations. The experiments showed that 5 runs were sufficient. The decision, which number K is best, has been taken by following a majority vote over all $K = 1, \dots, 10$ and for 5 different starting initialisations.

A continuous time-series is generating a continuous trajectory in phase space, which is a very difficult precondition for unsupervised clustering. Clustering with Multi-SOMs, and Multi-Neural-Gas performed well in the 3 dimensional case ($x(t), \dot{x}(t), \ddot{x}(t)$) even under this conditions. Although the results with the phase space alone were promising, we propose to extend the phase space by those features that have been found by the feature selection. For demonstration purposes, only an assortment of the implemented features is regarded.

4.1 TEST I

The first time-series (Fig. 3(a)) is artificial, consisting of 25,000 data points. Magnitude and duration of the peaks are random in a small range and the signal is contaminated with noise.

Feature Selection After determining the feature-series, calculating the correlation and covariance matrix and the values of every feature with respect to eq. (1), we got Tab. 1. The fact, that first and second derivative are

Feature	f_i	\dot{x}	\ddot{x}	FZC	NZC	TDZC	Mean	Var	Int	SC	SF	SR	RMS
Relevance	r_i	26	28	12	20	26	22	21	18	14	14	11	23

Table 1: Selection by Feature Relevance. Selected features are marked bold.

orthogonal to each other and are therefore the less correlated to other features, gives a good justification for our proposed approach to use the phase space as basis.

Feature Space Therefore, the phase space vector is extended by the following features: Fourth component is *TDZC*, fifth is *RMS* and last selected is *Mean*. The selected feature space is therefore of dimension $d = 6$:

$$\mathcal{F}^* = (x, \dot{x}, \ddot{x}, \text{TDZC}, \text{RMS}, \text{Mean})$$

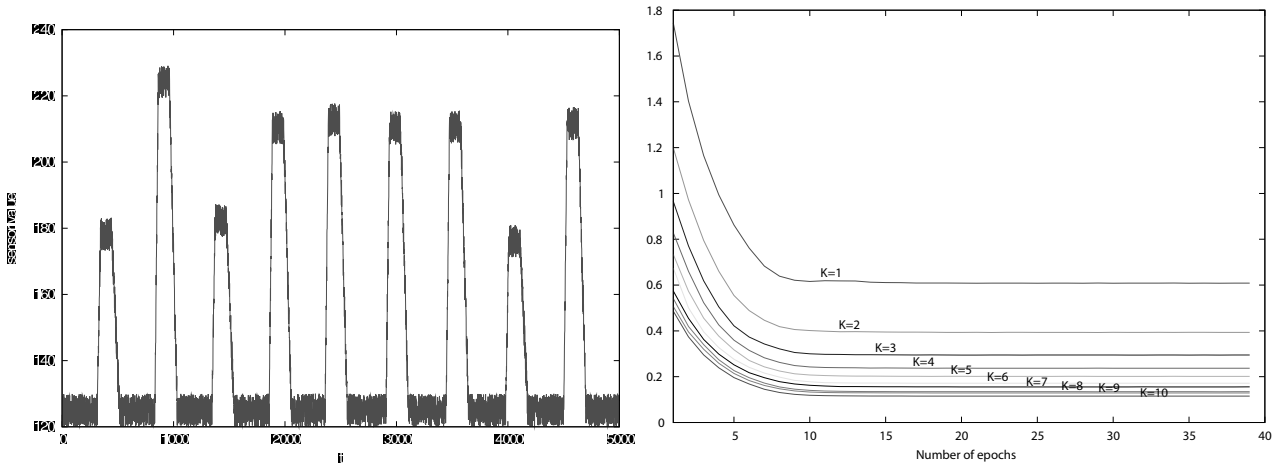


Figure 3: (a) Artificial time-series, first 5,000 data points. (b) Mean Deviation during learning for $K = 1, \dots, 10$. Even learning for just 15 epochs would have been sufficient.

Clustering with Multi-NGas of $K - 25$ type for a fixed K is done 5 times each with different initial starting positions. The results of the quality measures have been evaluated for every $K = 1, \dots, 10$ to yield an optimal number of clusters. The following table 2 shows the results of the Gap Statistic and Silhouette Coefficient:

k	1	2	3	4	5	6	7	8	9	10
Gap Statistics	0	0	0	1	0	4	0	0	0	0
Silhouette Coefficient	0	3	2	0	0	0	0	0	0	0
Maximum Value						4				

Table 2: Results of Gap Statistic and Silhouette Coefficient.

*Determining K^** Following Gap Statistics, the number of clusters is estimated in 4 out of 5 runs to 6. In this series the best K is 6. The (extended) Gap Statistic tends to a more detailed and sophisticated behaviour. With $K = 6$ the Multi-Neural-Gas is directly delivering a set of 6 clusters. The data points can be classified according to their winning neural gas. Evaluating this task for time-series yields a separation of the entire input space. The assigned classes of the original time-series can be identified with a colour (see Fig. 5 (a)).

Find representative The corresponding positions of the time-series for the 6 representatives are depicted in Fig. 6 (a) and have been marked with arrows. To determine the representative data point for each class, we propose to take that neuron, that maximises the quantity being the winning neuron for a data point during the last epoch.

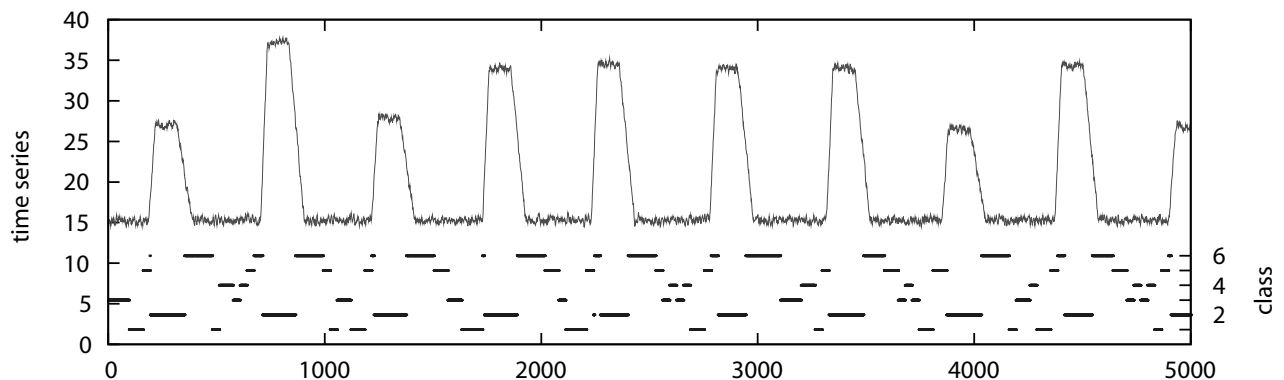


Figure 4: The original input time-series has been classified into six classes (depicted in lower part) by a $K = 6$, $N = 25$ Multi-Neural-Gas in 6 dimensional feature space.

Labeling of data means characterising and assigning real “events” or names to the found patterns. In Fig. 6(a) every class is associated with a number and colour. Pattern 6 can be characterised as the rising and falling edges, indicating in the second case the event of noise at the bottom. Bottom noise is being separated into three classes, namely 1, 3 and 4. Class number 5 occurs before curve rising. Pattern 2 could be interpreted as top of the hill.

4.2 TEST II

The second evaluated time-series (Fig. 5, depicted in upper part) was produced by an autonomous robot, consists of 2663 data points and corresponds to a short run of 26 sec. The entire real time-series is depicted in Fig. 5.

Feature Selection The corresponding feature-series were evaluated with respect to eq. (1) yielding:

Feature	f_i	\dot{x}	\ddot{x}	FZC	NZC	TDZC	Mean	Var	Int	SC	SF	SR	RMS
Relevance	r_i	26	28	8	15	9	9	10	5	7	8	7	7

Table 3: Selection by Feature Relevance. Selected features are marked bold.

Feature Space This result is quite similar to the one before (see Tab. 1). Again, the first and second derivative are the most linear independent features. The heuristically selected feature space is therefore 7 dimensional:

$$\mathcal{F}^* = (x, \dot{x}, \ddot{x}, \text{NZC}, \text{TDZC}, \text{Mean}, \text{Var})$$

*Clustering/Determining K^** The evaluation of clustering with Multi-NGas and of the quality measures Gap Statistics and Silhouette Coefficients (see Tab. 4) results in the most definite decision $K = 8$ proposed by the Silhouette Coefficient.

	k	1	2	3	4	5	6	7	8	9	10
Gap Statistics		0	0	0	1	1	1	2	0	0	0
Silhouette Coefficient		0	0	0	0	0	1	1	3	0	0
Maximum Value									3		

Table 4: Results of Gap Statistic and Silhouette Coefficient.

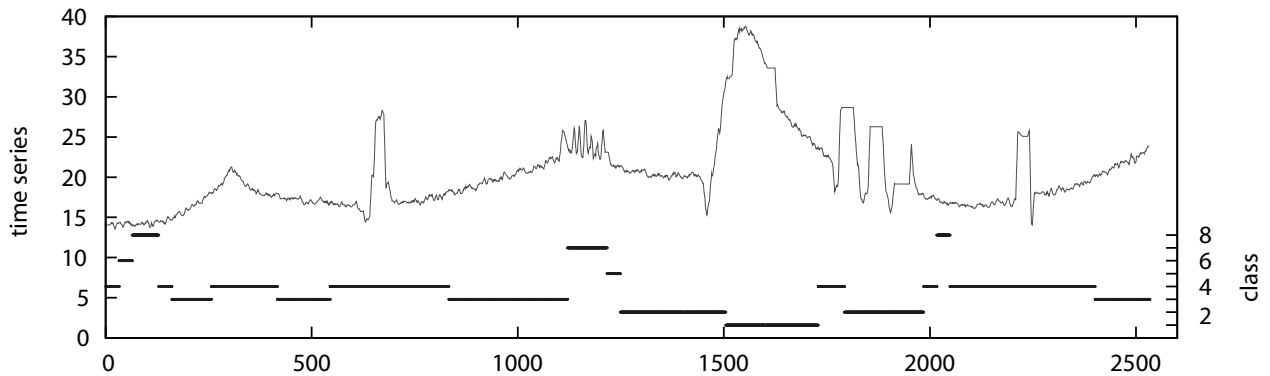
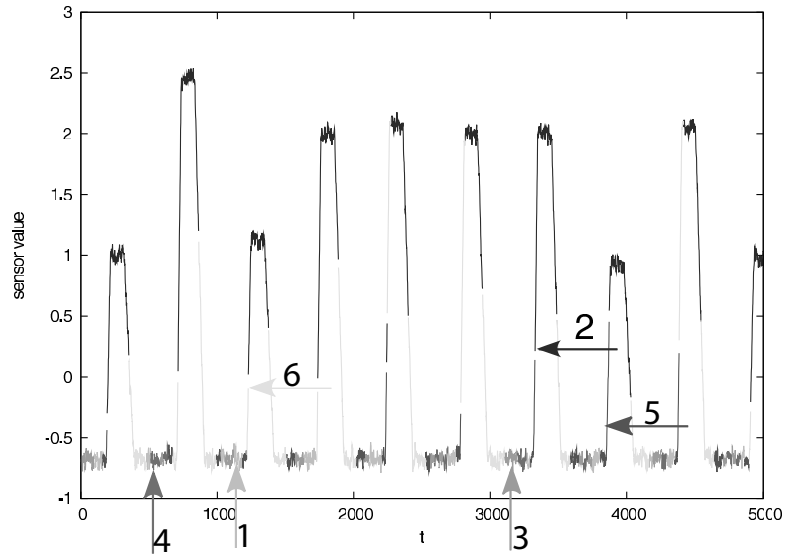


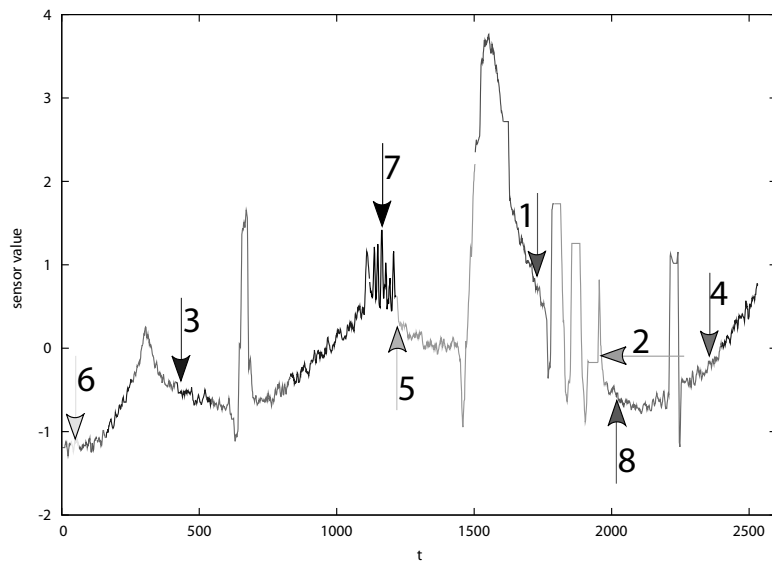
Figure 5: The original input time-series has been classified into eight classes (depicted in lower part) by a $K = 8$, $N = 25$ Multi-Neural-Gas in 7 dimensional space.

Find representative The data points can be classified according to their winning neural gas. The positions of the 8 representatives for the second time-series are depicted in Fig. 6 (b) and have been marked with arrows.

Labeling means assigning real events to found patterns. Though excluding excessive details of the underlying environment the time-series was recorded in, the five different types of the found patterns seem plausible to a human observer.



(a)



(b)

Figure 6: The corresponding 6 and 8 classes respectively of the time-series are marked with different colours. Additionally, the representatives of each class found are marked with arrows. All found classes seem plausible to a human observer.

5 CONCLUSIONS

We propose a system for Pattern Discovery in Time-Series by unsupervised feature selection and unsupervised, self-organised clustering. It has been shown to be a valuable framework even under the assumption that no prior knowledge about the time-series itself and suitable derived features is available. There is no additional information required extra to the time-series.

The first task accounted for is Feature Selection. The proposed feature selection algorithm is determining the feature relevance for a variety of transformations. We propose to take the contiguous phase space as a method of nonlinear dynamics as basis and extend it by the selected features as extra dimensions. No search or clustering in advance is needed and thus, determining of feature space is fast.

The Multi-SOMs and Multi-Neural-Gas have been demonstrated to be useful neural network tools for unsupervised and self organising clustering. The extended cluster analysis coefficients are capable of judging the quality of reached clustering and estimating the optimal number of partner networks.

The presented approach for discovering patterns has been evaluated with time-series from a robotic application with promising and plausible results, even under the assumption of simple majority decisions for A) Feature Selection and B) Estimation of the optimal number of clusters.

Further investigation is needed to make the whole system learn the selection of features from the results of quality measures after clustering by Reinforcement Learning or Evolutionary Algorithms.

REFERENCES

1. Goerke, N.; Kintzler, F.; Eckmiller, R.: "Self Organized Partitioning of Chaotic Attractors for Control", In: Proc. of the Int. Conf. on Artificial Neural Networks (ICANN'01), Springer, LNCS 2130, pp. 851-856, (2001).
2. Goerke, N.; Kintzler, F.; Brüggemann, B.: "Multi-SOMs for Classification", In: Proc. of the Int. Workshop on Automatic Learning and Real-Time, ALaRT'05, pp. 99-107, (2005).
3. Goerke, N.; Scherbart, A.: "Classification using Multi-SOMs and Multi-Neural Gas", In: Proc. of IEEE World Congress on Computational Intelligence, (WCCI 2006).
4. Goerke, N.; Scherbart, A.: "Using Multi-SOMs and Multi-Neural-Gas as Neural Classifiers", In: Proc. of Ind. Conf. on Data Mining (ICDM06), Springer, LNCS 4065, pp. 250-263, (2006).
5. Haykin, S.: "Neural Networks: A Comprehensive Foundation", Prentice Hall, Erlangen (1999).
6. Kaufman, L.; P.J. Rousseeuw, P.J.: "Finding Groups in Data: An Introduction to Cluster Analysis", Wiley, New York, (1990).
7. Kohonen, T.: "Self-Organized Formation of Topologically Correct Feature Maps", *Biological Cybernetics*, Vol. 43, pp.59-69, 1982.
8. Kohonen, T.: "Self-Organizing Maps", Springer, Berlin Heidelberg 1995.
9. Martinetz, Th.; Schulten, K.: "Neural-Gas Network Learns Topologies", *Artificial Neural Networks*, Vol. I, pp. 397-402, (1991).
10. Martinetz, Th.; Berkovich, S. G.; Schulten, K.: "Neural Gas: Network for Vector Quantization and its Application to Time-Series Prediction", IEEE 1993.
11. Mitra, P.; Murthy, C.A.; Pal, Sankar K.: "Unsupervised Feature Selection Using Feature Similarity", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 4, April 2002.
12. Mörchen, F.: "Time series feature extraction for data mining using DWT and DFT", Techn. Rep. No. 33, Dept. of Mathematics and Computer Science, Philipps-University Marburg, 2003.
13. Strogatz, S.: "Nonlinear Dynamics and Chaos With Applications To Physics, Biology, Chemistry, And Engineering", "Westview Press", "1994".
14. Tibshirani, R.; Walther, G.; Hastie, T.: "Estimating the Number of Clusters in a Dataset via the Gap Statistics", Tech. Rep. 208, Dept. of Statistics, Stanford University, 2000.
15. Tzanetakis, G.; Cook, P.: "Musical genre classification of audio signals", *IEEE Trans. Speech Audio Processing*, vol. 10, no 5, p. 293-302, 2000.
16. Zell, A.: "Simulation Neuronaler Netze", Oldenbourg Verlag, 2000.