Pattern Recognition in Metabonomics Using Self Organizing Maps

Stefan W. Roeder[1], Ulrike Rolle-Kampczyk[1], Olf Herbarth[1,2]
[1]UFZ – Centre for Environmental Research Leipzig – Halle Ltd.
Department of Human Exposure Research and Epidemiology
P.O.Box 500135, 04301 Leipzig, Germany
stefan.roeder@ufz.de
[2]Faculty of Medicine, University of Leipzig,
Liebigstrasse 24, 04103 Leipzig, Germany

ABSTRACT

The contiguity between external exposure and internal stress burden in humans is linked by metabolic pathways. To gather deeper knowledge about these links it is necessary to look behind the scenes and to find out which variables affect each other as well as which groups of individuals can be described.
The approach shown in this paper is based on the algorithm of self organizing maps (SOM), which have shown their capabilities for clustering tasks previously. They enable us to subdivide cases into classes without prior knowledge. Furthermore prediction of selected outcome variables is possible after training of the algorithm. New visualization techniques atop of the SOM algorithm allow a multidimensional view on the data and their immanent structures as well as on the classes found.
Finally we checked several parameter combinations of the SOM for their applicability and performance with our data.

Keywords: stress patterns, pattern recognition, artificial neural networks, self organizing maps, metabonomics

1. INTRODUCTION

To gather deeper knowledge about human diseases caused by environmental influences it is necessary to associate external and internal stress patterns as well as binary outcome variables about the health state of each individual. The palette of external stress patterns embraces environmental factors as well as socio-economic factors.

This paper describes an approach for associating external and internal stress patterns in order to recognize patterns in these combined patterns. These patterns are used to deduce on the relationships and processes behind.

The main focus of this investigation is to classify cases into several typical clusters and subsequently to find core features of these clusters.

2. MATERIAL AND METHODS

Internal stress patterns are described by metabolites which are generated during metabolism. They can be measured via Liquid Chromatography/Mass Spectrometry (LC-MS/MS). Analysis of spectroscopic patterns with multivariate statistical methods is often referenced as „Metabonomics“. Although LC-MS/MS patterns contain a lot of useful information about the current state of the organism of origin, it is very difficult to handle these extremely large data sets with traditional statistical methods. The use of principal component analysis (PCA) in combination with NMR-spectra is very frequent [1]. However, these methods assume a linear relationship between the measured sample and the converted principal component. Because in biological processes this linear relationship can not be assumed, errors are not controllable.

Our approach is different in several ways: In difference to Ebbels et al. who uses rodents for its research our subjects are humans. With rodents one is able to vary environmental variables. They apply compounds orally to rats and measure the stress burden from their urine. This approach is not applicable to humans because

a) lab conditions have to be ensured,
b) a control group must be available,
c) the concentrations applicable are larger than applicable to humans.

With humans such experiments are not possible. We are unable to experiment with them. The only way is to measure the values of everyday life and to look for differences.

Other approaches compare structured stress spectra [2]. A restriction is that these spectra are equally structured. This is not suitable for our situation because stress spectra of outer and inner burden are not of the same structure.

Therefore we decided to use a clustering method based on self organizing maps (SOM) [3;4] . They provide an efficient way to map from an n-dimensional space to a two-dimensional space and to visualize multivariate data. The self organizing map is a member of the class of the unsupervised artificial neural network algorithms. In contrast to supervised artificial neural network algorithms, self organizing maps are able to extract typical features without any prior knowledge about the structure of the data.

Comparable approaches are very rare. We found only one paper which deals with the use of self organizing maps in metabonomics [5].

2.1 Algorithm of Self organizing map

The main advantage of self organizing maps is their ability to map a high-dimensional data set onto a lower dimensional (usually two-dimensional) space while preserving the original topological relationship between the objects in the data set. The self organizing map consists of two layers: the input layer and the Kohonen layer. Both layers are fully interconnected. The input vectors x_i and the weight vectors w_i have the same number of dimensions, which is equal to the number of variables inside the data set under consideration.

The learning algorithm of the self organizing map is iterative. In a first step all weight vectors of the Kohonen layer are initialized with random values. During each iteration a single input neuron is randomly selected. The distance between this neuron and each neuron in the Kohonen layer is then calculated. Different metrics can be used [6]. Assuming our data is scaled metric the Euclidean distance de fits best:

$$d_e = \sqrt{\sum_{i=1}^{p} (x_i - y_j)^2}$$

The Kohonen neuron with the least Euclidean distance to the selected input neuron is selected as winner neuron. Its weight vector is moved towards the weight vector of the selected input neuron in Euclidean space using the following formula:

$$w_i(k+1) = w_i(k) + \alpha * e^{(-d_e^2 / 2*\sigma^2)} * (x_i - w_i(k))$$

The term $e^{(-d_e^2 / 2*\sigma^2)}$ describes the size of the neighbourhood around the winning neuron in the Kohonen layer. Farther neurons are less affected by weight changes than nearer neurons.

The lattice type of the Kohonen layer can be taken as rectangular or hexagonal [7]. We used rectangular type for our analysis. Parameters α and σ are monotonically decreasing by multiplying them with a factor below 1 called momentum, leading to convergence of the Kohonen layer after training. This process is iterated until a predefined number of cycles and a stable state is reached. Details of the algorithm can be found in [7] for theoretical considerations.

One drawback of the self organizing map is that the algorithm has no benefit on parallel computers [8]. To elect the winning neuron, the algorithm has to check the distance to each neuron in the Kohonen layer which requires at least n/2 comparisons. On the other hand side the self organizing map provides the advantages of direct access to the stored knowledge, which makes it possible to predict new cases by applying SOM learning rules to them. The properties of the winning Kohonen neuron can be treated as core properties of the respective case.

2.2 Pattern recognition

In our approach we observe environmental factors from indoor air, socio-economic situation as well as nutrition variables. In indoor air samples we measure volatile organic compounds (VOC). Information about socio-economic situation and nutrition is gathered by a structured questionnaire. Internal stress patterns are described in terms of mass spectra. They were gathered by LC-MS/MS analysis of urine samples.

Both stress patterns are described by a data vector for each individual case. After combining both vectors each case is represented by a data vector describing its situation (see figure 1). Combining is a simple casewise concatenate of the variables. Sequence is not important, it has only to be ensured that sequence is the same in all cases.

All variable values are normalized to a closed {0;1}-interval before processing [8]. This is done by the following assignment:

$$x_{ij}norm = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)}$$

with i as the variable index and j as the case index. These combined data vectors are fed into a self organizing map.

The number of neurons in the Kohonen layer predicts the maximum number of clusters simultaneously.

After the training steps each Kohonen neuron represents the centre of a cluster. Each cluster contains a typical set of cases with typical properties. These properties are best represented by the data vector of the Kohonen neuron in the center of each cluster. Therefore the values of the data vectors of each Kohonen neuron can be taken into account as best describing the environmental stress situation in relationship to the outcome situation found in the cases of this cluster.

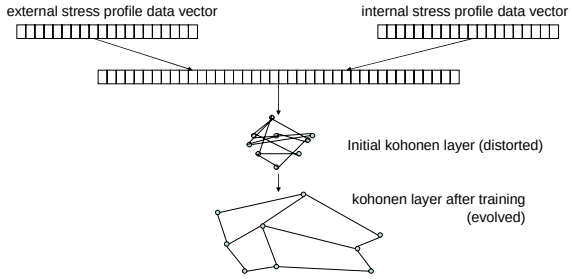

figure 1 vector matching and learning

The two-dimensional graphical representation we are using in our tool implementing the Kohonen layer is able to map each distinct vector element (dimension) from the input vectors to either x- or y-axis. Using this technique it is possible to view each dimension also in a n-dimensional space. One can show the relationships between each selected variable pair.

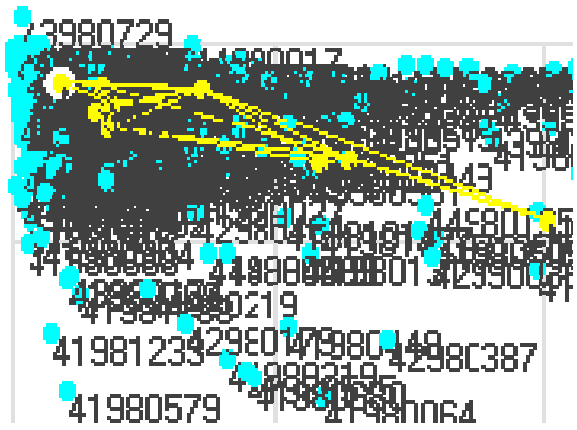

figure 2 graphical SOM representation

Each dot represents a case with its assigned case identification number. The dots which are connected by the line represent the Kohonen layers graphically. In this case it is a 3x3 network.

2.3 Classification

As shown in the previous paragraph, SOM are only capable to map from an n-dimensional attribute space to a lower dimensional (typically 2-dimensional) attribute space. To find class assignments it is necessary to add some calculations.

Class assignment for each case can be determined by calculating the Euclidean distance of the cases weight vector against the weight vector of each Kohonen neuron after training using the following equation.

$$d_e = \sqrt{\sum_{i=1}^{p} (x_i - y_j)^2}$$

X and Y represent the weight vectors for measuring the distance in between. P stands for the number of dimensions of both vectors.

The Kohonen neuron with the least Euclidean distance is the centre neuron of the class, the case in doubt belongs to. One can count the number of assigned cases for each Kohonen neuron. Kohonen neurons with large numbers of assigned cases represent class centres. The graphical representation of this dependency is shown in figure 3. This approach is a modified version of the visualization of the U-matrix first described by Ultsch [9], who uses a two-dimensional visualization and colour coding of data density.

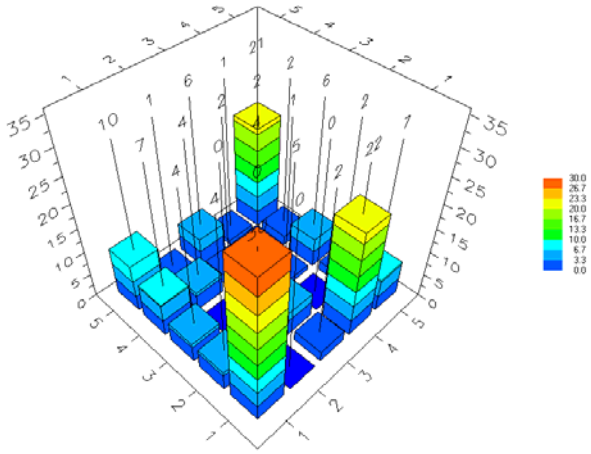

figure 3 graphical representation of case assignment for a 5x5 Kohonen layer

The topological position of the neuron on the map is given by the numbers on x- and y-axis. The height of

each bar is equal to the number of cases which are assigned to this neuron. Additionally the case-identifier and variable values behind each neuron can be obtained.

During several learning cycles a selected case can be assigned to different neurons. This is a result of the random initialized weight vectors in the Kohonen layer and the subsequent evolving process. If the same group of cases is assigned to different neurons this is no problem at all. It is also no big deal if several cases jump from one to another group. Difficulties arise if the group members change rapidly. Measuring of the permanence of class assignment of a single case over several learning cycles is necessary, but remains a task for future development.

Comparing clustering solutions generated by different SOM topologies with possibly different class numbers can be done by applying Davies Bouldin Index [10].

2.4 Calculation of parameter values for Kohonen neurons

For the analysis of the results of clustering it is important to gather deeper knowledge about the variable values behind each Kohonen neuron as well as the relation of these values between all Kohonen neurons. Because of the fact that the weight vectors of the Kohonen neurons contain normalized values in a closed {0;1}-interval these values must 'denormalized' be before considering on them. Denormalization means the opposite transformation than normalization and transforms the values of the weight vector in a Kohonen neuron back to the original co-domain of the underlying variable using the following equation:

$$x_{ij} = (x_{ij} norm * (\max(x_j) - \min(x_j)) + \min(x_j)$$

We use a visualization scheme for this purpose, which shows the values of a variable for all Kohonen neurons simultaneously.

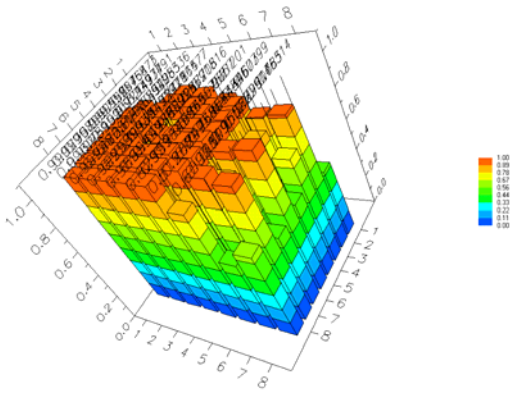

figure 4 visualization of values of a specific variable inside a 8x8 Kohonen layer

In figure 4 each Kohonen neuron is visualized by a bar, which is color coded and labeled with the denormalized value of the selected variable. Using these denormalized values the distribution of values can be inspected for each variable. The values behind Kohonen neurons which are not assigned as class centers are unimportant and can be removed from further inspection. Values behind Kohonen neurons representing class centers are at the same time the approximated values for the respective variable in this class.

2.5 Sensitivity against parameter changes

SOM are affected by changes in parameter settings. This is the drawback of this approach: many parameters have to be predefined for proper function, but nobody knows the optimal settings. They can only be obtained by experimenting with the settings. This is far away from unprejudiced research.

Therefore we had to investigate the sensitivity against these changes. Initially, the parameters α and σ and momentum have been selected as follows:

α = 0.1

σ = 2

momentum = 0.999.

We stopped the training after 2000 iterations. The success of training was visually inspected in terms of the evolvement of the Kohonen layer. Complete evolvement is necessary for reasonable mapping performance. Evolvement is difficult to measure. The following figure gives an example. The left part shows a 2-dimensional data set (dots) with a distorted 3x3 Kohonen layer above, whereas the right side shows the same data set with a completely evolved 3x3 Kohonen layer.

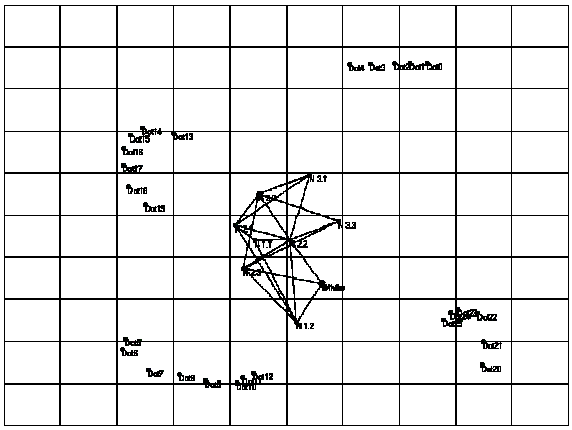

figure 5 distorted 3x3 Kohonen layer above an artificial 2-dimensional data set

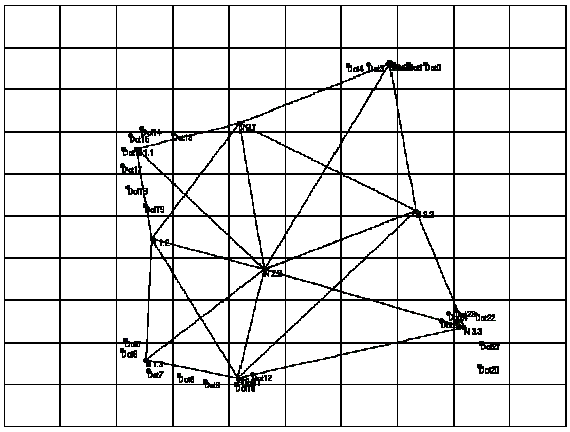

figure 6 evolved 3x3 Kohonen layer above an artificial 2-dimensional data set

Because of the lack of measures for map evolvement, we inspect the Kohonen layer visually and decide whether it is distorted or not.

For better understanding of the sensitivity of the algorithm against parameter settings we tested several parameter settings. Variations were done in the learning rate α, the neighbourhood σ and momentum.

α	σ	momentum	result
1	2	0,999	correct evolvement; convergence
0,5	2	0,999	correct evolvement; no convergence
0,5	1,8	0,999	correct evolvement; no convergence
1	1,5	0,999	Bad evolvement; converges too fast
0,5	1,5	0,999	no evolvement
2	1,8	0,999	no evolvement
1	1,8	0,999	bad evolvement
0,8	1,8	0,999	correct evolvement; no convergence
1	1	0,999	no evolvement
0,5	0,8	0,999	no evolvement

table 1 sensitivity of parameter changes on a 3x3 Kohonen layer above an artificial 2-dimensional data set

Correct evolvement means that the 3x3 Kohonen map evolves to a nearly rectangular shape with not more than one crossing between the connecting lines of the neurons. Convergence describes that the map comes to a stable state with rising number of learning cycles. Optimal parameter settings unite topological correct evolvement with convergence.

We found best settings for a 3x3 Kohonen layer with α =1 the neighbourhood σ =2and momentum=0.999.

Another variable factor is the number of neurons in the Kohonen Layer of the SOM. Usually the number of dimensions for the Kohonen layer should be equal to the number of dimensions in the input layer. The number of neurons in each dimension is variable. A approximation of these values is possible. As a rule of thumb one can assume that the number of neurons should equal or greater than the number of classes to be found.

Céréghino et al. assume that the should not be too large, because it will then “overfit” the models. Optimum size was determined by training of different sized SOM and comparing them in terms of minimum quantization error. Quantization error is measured as the distance between each input vector and its nearest neuron in the Kohonen layer [11].

On the other hand side if the map is too small there is no chance to recognize outliers. Typically, an outlier takes “it’s own” neuron in the Kohonen layer, if enough neurons are present. Therefore it cannot affect other neurons in the Kohonen layer by distorting them.

The correct number of neurons in the Kohonen layer depends in the data set used and its inherent classes. It should be approximated in several steps by measuring the SOM quality using quantization error and Davies Bouldin Index.

2.6 Prediction of outcome variables

Another area for application SOM is the prediction of outcome variables. This is very close to classification, because it will enable us to predict future health state of study participants from simply a urine sample. For prediction of outcome variables we trained the self organizing map with cases which outcomes were known. In a second step the unknown cases for prediction were fed into the map. The Euclidean distance against each Kohonen neuron is calculated as described above. The Kohonen neuron with the least distance represents the best matching scenario. Its value for the outcome variable under consideration is the most probable future value for the selected case.

For evaluation purposes the calculation of a correct prediction rate is essential. This can be done by dividing a given data set into a training set and a test set which are nearly equal sized. The training set is used to train the SOM to a stable state. Afterwards the test set is used to verify the outcomes prognoed by the SOM. During the forecasting process it is important not to use the outcome variables for calculation of the euclidean distance. This means with n variables (dimensions) and k outcome variables to predict, the euclidean distance is only calculated from the n-k remaining variables.

3. RESULTS

Unfortunately, metabonomics is in a very early stage of development and therefore many of the necessary tools are not available yet [12]. Because of this lack we built up our own tool for data handling, preparation, SOM training and visualization. The approach presented here integrates data vector normalization, SOM training, class finding and variable prediction.

In a first step we were able to determine settings for SOM parameters, which guided us to a stable configuration. We were able to clearly see the fact, that the SOM algorithm does not force all neurons to be populated. Consequently there is an upper limit for the number of neurons in each dimension of the SOM. We examined this limit by visual inspection and gradually lowering the number of neurons in each dimension of the SOM.

If the number of neurons is too high then a significant number of neurons remain unassigned. These unassigned neurons can be perceived as borders between the classes. Simultaneously the number of assigned input vectors to each assigned neuron is relatively low. Typically one can observe several contiguous Kohonen neurons with only slight differences. By reducing the number of Kohonen neurons these minor differing neurons will be summarized into one neuron representing the respective class. This process of reconfiguration and visual inspection is supported by indicators for classification quality such as Davies-Bouldin-Index. Future planning includes automation of this process using a computer grid.

Using this approach we are able to classify data sets with large number of variables into typical clusters. In terms of metabonomics this will be used to establish links between internal stress patterns and external exposure to gather clinical relevant information for practitioners.

In consequence the shown approach is easy to handle and does not require complex equipment for analysis. It can therefore be established in nearly every laboratory provided with state-of-the-art computatinal equipment.

4. DISCUSSION

Using this method we are able to
a) recognize dependencies between external and internal stress situation as well as
b) predict the probabilistic value of an outcome variable for a given environmental stress situation.

Examinations with data from different studies realized at our department [13;14] have shown the usefulness of this approach. Future plannings include the application of this method to data from longitudinal studies from birth cohorts to gain knowledge about critical time windows in childhood development.

REFERENCES

[1] Ebbels T, Keun H, Beckonert O, Antti H, Bollard M, Holmes E, Lindon J, & Nicholson JK, "Toxicity classification from metabonomic data using a density superposition approach: 'CLOUDS'," *Analytica Chimica Acta*, 490(2003), pp. 109-122, 2003.

[2] J. E. Katz, D. S. Dumlao, S. Clarke, & J. Hau, "A new technique (COMSPARI) to facilitate the identification of minor compounds in complex mixtures by GC/MS and LC/MS: tools for the visualization of matched datasets," *Journal of the American Society for Mass Spectrometry*, 15(4), pp. 580-584, 2004.

[3] Kohonen T, *Self Organizing Maps*, Series in Computer Sciences ed. Heidelberg: Springer, 1997.

[4] Zampighi LM, Kavanau CL, & Zampighi CA, "The Kohonen self-organizing map: a tool for the clustering and alignment of single particles imaged using random conical tilt," *Journal of Structural Biology*, 146(2004), pp. 368-380, 2005.

[5] L. K. Dow, S. Kalelkar, & E. R. Dow, "Self-organizing maps for the analysis of NMR spectra," *Drug Discovery Today: BIOSILICO*, 2(4), pp. 157-163, 2004.

[6] A. Zell, "Simulation neuronaler Netze," München: Oldenbourg, 2000.

[7] Kohonen T, "Self-Organizing Maps,", 3rd ed New York: Springer, 2000.

[8] Zupan J and Gasteiger J, "Neural Networks in Chemistry and Drug Design," Weinheim: Wiley-VCH, 1999.

[9] A. Ultsch, "Self-organizing Neural Networks for Visualization and Classification," in *Information and Classification*. O. Opitz, B. Lausen, and R. Klar, Eds. Berlin: Springer, 1993, pp. 307-313.

[10] Davies DL & Bouldin DW, "A cluster separation measure," *IEEE Trans. Pattern Anal. Machine Intell.*, 1(4), pp. 224-227, 1979.

[11] R. Cereghino, F. Santoul, A. Compin, & S. Mastrorillo, "Using self-organizing maps to investigate spatial patterns of non-native species," *Biological Conservation*, 125(4), pp. 459-465, 2005.

[12] L. W. Sumner, R. A. Dixon, & P. Mendes, "Plant metabolomics: Large-scale phytochemistry in the functional genomics era," *Phytochemistry*, 62(6), pp. 817-836, 2003.

[13] M. Borte, Schulz R., I. Lehmann, and etal., "Influence of lifestyle and behaviour on the development of the immune system and allergic diseases. The LISA birth cohort study," in *Public Health Research and Practise*. Merker N, Göpfert P, and Kirch W, Eds. Regensburg: S. Roderer Verlag, 2001, pp. 59-77.

[14] U. Diez, M. Rehwagen, U. Rolle-Kampczyk, H. Wetzig, R. Schulz, M. Richter, I. Lehmann, & M. Borte, "Redecoration of apartments promotes obstructive bronchitis in atopy risk infants - Results of the LARS study," *Int J Hyg Environ Health*, 206, pp. 173-179, 2003.