Self-organizing maps and Sammon’s mapping in environmental informatics and bioinformatics

Mikko Kolehmainen¹, Teri Hiltunen¹ and Harri Niska¹

Abstract
The aim of this study was to evaluate the usability of self-organizing maps (SOM) and Sammon’s mapping in analysing data originating from processes of environmental informatics and bioinformatics. The methods used are shortly introduced. The results show that the combination of SOM and Sammon’s mapping has great potential in data exploration, and can be used to reveal important features of the measurement techniques used, reveal new information about phenomena studied earlier, speed up research work at laboratory level, act as a hypothesis generator for directing traditional research, and supply clear and intuitive visualization of the environmental phenomenon studied.

1. Introduction
The problem encountered in today’s world is that of information overflow. Therefore, the enrichment process turning the measured data into information and knowledge is a very important subject of study. This was the starting point of a series of case studies with focus on environmental and bioprocesses.

An important trend during the last decades has been the advancement of medicine and biochemistry. This has culminated to micro-array technology, which enables to monitor the working of the genes. This opens new prospects also in environmental sciences in the form of monitoring of environmental processes at a new level of accuracy.

A new brand of computation called computational intelligence (CI) has been developed during the last two decades. Methods of CI create a data processing system by utilizing measurements of the study subjects and then constructing more abstract objects algorithmically mimicking the processes found in the brain (e.g. neural net-
works) or in the nature (e.g. evolutionary computation). The methods of CI have also been applied successfully in many industrial and everyday applications.

The aim of the study was to evaluate self-organizing maps and Sammon’s mapping for analysing environmental and bioinformatics problems, using a number of selected case studies. Air quality and bioinformatics were selected as domains of application because these areas had suitable data available, as well as experts in the field who could participate in guiding the work.

The main innovation of the study carried out was to show how the performance of SOMs can be enhanced by combining them with Sammon’s mapping. This idea has been presented earlier (e.g. Chang and Lee, 1973), but it has not been previously studied to this extent with real world problems and datasets. Additionally, the Visual Data software (www.visipoint.fi) developed and tested in these studies takes the approach to a practical level, where end-users who are not familiar with the computational methods can produce results in their own studies (see Figure 1.). The study described here has also been reported in doctoral dissertation by Kolehmainen (2004).
2. Materials and methods

2.1 Self-organizing map

The self-organizing map (SOM) is one of the best known unsupervised learning methods (Kohonen, 1997). It can be summarized as follows.

1. Find the Best-Matching Unit (BMU) for one input vector according to the minimum Euclidean distance.

2. Move the weight vector of the BMU towards that input vector, using the update rule.

3. Move the weight vectors of neighbouring neurons (according to the neighbourhood function) towards that input vector, using the update rule.

4. Repeat steps 1-3 for the next input vector until all input vectors have been used.

5. Repeat steps 1-4 until convergence.

6. Find the final BMU (the neuron which the individual belongs to) for each input vector according to the Euclidean distance.

A variation of the SOM, the tree-structured SOM, was used in the work (Koikkalainen, 1994). The software implementation thus consists of several SOMs organized hierarchically in several layers in a pyramid-like fashion. The number of neurons at a lower level is four times the number at the previous level. However, visual inspection of the measurement data can be directed to one level at a time and it can thus be used for data exploration similarly to a SOM created with the original algorithm. The learning rule of the TS-SOM takes a simpler form and is described in Equation 1 as follows.

\[
\mathbf{w}_m(t+1) = \mathbf{w}_m(t) + \alpha(t)[\mathbf{x}_i - \mathbf{w}_m(t)]
\]  

where \(t\) is a counter for iterations, \(\mathbf{x}_i\) is the input vector, \(\mathbf{w}\) is the weight vector of the BMU and \(m\) is the index for it.

In TS-SOM the neighbourhood function has been reduced to a fixed form where the neighbourhood is always defined to be the four adjacent neurons of the BMU. This is possible since the role of the altering neighbourhood radius is handled by more coarse levels of the TS-SOM structure.

2.2 Sammon’s mapping

Sammon’s mapping (Sammon Jr, 1969) is a non-linear mapping algorithm closely related to the metric version of multi-dimensional scaling (MDS) (Torgerson, 1952). The aim of the Sammon’s mapping algorithm is to represent points of \(p\)-dimensional space in 2-dimensions. The algorithm adjusts the locations in the 2-dimensional tar-
get space so that the original structure of the measurement vectors in p-dimensional space is conserved maximally.

Sammon’s mapping represents the relative distances of vectors in a measurement space and is therefore useful in determining the shapes of the clusters and the relative distances between them. Even though the SOM and Sammon’s mapping work quite similarly, the numerical calculation of the latter is more time consuming with exponential growth as a function data points. Because this is usually a problem with a massive data set, it is of benefit to combine these two algorithms, which was consequently the starting point of this study.

2.3 U-matrix representation

The U-matrix representation (Ultsch and Siemon, 1990) is a way of describing the relative distances between the neurons in a SOM map. It is calculated by taking the mean of the distances between the p-dimensional weight vectors of neighbouring neurons of the target neuron. The result of the calculation is a number assigned to each neuron on the map. These numbers can then be used as grey level values or as a Z-axis dimension in visualization (see Figure 2).

2.4 Application of the methods

Self-organizing maps combined with Sammon’s mapping were applied to several datasets originating from time-series of air quality measurements, time-series of yeast fermentation process, gene expression data and epidemiological data. In all these case studies, the strong visualization properties of the previously mentioned methods were used to reveal the cluster structure in the data and to explain the meaning of the phenomena found in this way.

3. Results and discussion

In the first case study (Kolehmainen et al., 2001), the data exploration procedure by combining the SOM and Sammon’s mapping was applied to Ion Mobility Spectrometry (IMS) data of odorous sulphuric compounds measured from a pulp mill chip-feed process. The combined method was shown to reveal important features of the measurement technique itself, e.g. the separability of different sulphur compounds, and their concentrations, as well as provide a clear and intuitive view of the complex and multivariate problem domain of IMS based measurement in general.
The same measurement technique (IMS) and combined computation method was also applied to monitor yeast fermentation process (Kolehmainen et al., 2003). The results showed that different phases of one fermentation batch and different batches can be distinguished by their distinct IMS profile (see Figure 2.). Moreover, the phases detected by this method and those found in standard textbooks differed substantially in the early phases of the process. More specifically, our method was able to detect two phases instead of the known one. This shows how an innovative application of the method described can reveal new information in a way that is novel to the whole discipline of bioprocess engineering, for example.

In the third case study the methods were applied to gene expression data (Törönen et al., 1999). We were able to show how these methods can be used to find
and identify the most interesting genes of a certain phenomenon from thousands of genes. Consequently, the use of such software in screening potential genes speeds up the research work at the laboratory level, enabling a faster cycle of experiments.

In the fourth case study epidemiological data of 1650 subjects (Valkonen et al., 2002) was explored. Using 25 biochemical and physiological variables potentially connected to insulin resistance syndrome, we showed how the inspection of multivariate data can lead to identification of important cluster structures consisting of healthy individuals, groups of individuals with severe symptoms, and two intermediate groups. It was thus demonstrated how this method can act as a hypothesis generator by finding new groupings and unidentified groups from multivariate research data. The SOM and Sammon’s mapping for this particular study are shown in Figure 3.

![Figure 3](image-url)

Self-organizing map (left) with bar graphs showing the distribution of the central variables: blood glucose (1st bar), serum insulin (2nd bar), triglycerides (3rd bar), HDL cholesterol (4th bar), systolic blood pressure (5th bar) and waist-to-hip ratio (6th bar). The clusters marked with c1-c4 have been determined using the corresponding Sammon’s mapping, which is shown on the right. (modified from Valkonen et al., 2002).

In the fifth case study (Kolehmainen et al., 2000) SOM and Sammon’s mapping were used to cluster air quality data. The results showed that data exploration can be used to yield clear and intuitive visualization about the air quality episodes (intervals of bad air quality) for the area inspected. Moreover, it was shown how the episodes
can be detected from the rest of the measurement data, identified using standard statistical indicators and assigned a textual description for later usage

4. Summary

The large potential of combining SOM and Sammon’s mapping was demonstrated in several studies applying the combined usage of the SOM and Sammon’s mapping. The results showed that such hybrid use can (i) reveal important features of the measurement technique itself, (ii) reveal new information in a way that is novel to the whole discipline, (iii) speed up the research work at the laboratory level, (iv) act as a hypothesis generator for more focused traditional research, and (v) supply clear and intuitive visualization about environmental phenomena, such as air quality episodes, for the area inspected. Overall, it can be concluded that if the method tested is implemented as user-friendly software (e.g. Visual Data), excellent results can be achieved.

Bibliography