ESANN'2009 proceedings, European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning. Bruges (Belgium), 22-24 April 2009, d-side publi., ISBN 2-930307-09-9.

The Exploration Machine – A Novel Method for Structure-Preserving Dimensionality Reduction

Axel Wismüller

Depts. of Radiology and Biomedical Engineering, University of Rochester, New York
601 Elmwood Avenue, Rochester, NY 14642-8648, U.S.A.

Abstract. We present a novel method for structure-preserving dimensionality reduction. The Exploration Machine (Exploratory Observation Machine, XOM) computes graphical representations of high-dimensional observations by a strategy of self-organized model adaptation. Although simple and computationally efficient, XOM enjoys a surprising flexibility to simultaneously contribute to several different domains of advanced machine learning, scientific data analysis, and visualization, such as structure-preserving dimensionality reduction and data clustering.

1 Motivation

The exceedingly growing amount of high-dimensional information stored in computer-accessible data bases and web resources raises the question of how to organize and extract useful knowledge from this abundant material. Here, a key issue is structure-preserving data reduction which has moved into the focus of interest as an issue of high-priority research efforts. Structure-preserving data reduction can frequently be accomplished by two alternative approaches, namely data partitioning in the sense of 'clustering' and dimensionality reduction often referred to as 'embedding'.

In this paper, we introduce a novel algorithm called 'Exploration Machine' (Exploratory Observation Machine – XOM) that unexpectedly can approach *both* domains from a unified viewpoint within a *single* computational framework. After explaining the XOM algorithm, we present applications to the analysis of multidimensional biomedical data from different real-world domains: whole-genome microarray gene expression experiments and functional MRI analysis. These two different examples convey XOM's capability to simultaneously contribute to the aforementioned different challenges of pattern recognition, namely structure-preserving dimensionality reduction *and* clustering.

2 Algorithm

The Exploration Machine (XOM) algorithm can be resolved into three steps. For simplicity, let us first consider N real-valued input vectors $\mathbf{r}_i$ in the 'observation space' O, each of dimensionality D.

Step 1: Define the topology of the input data in the observation space O by computing distances $d(\mathbf{r}_i, \mathbf{r}_j)$ between the data vectors $\mathbf{r}_i, i \in \{1, \ldots, N\}$. This step is omitted, if the input data is already given as a set of distances between

input data items.

Step 2: Define a 'hypothesis' on the structure of the data in the embedding space E, represented by 'sampling' vectors $\mathbf{x}_k \in E$, $k \in \{1, \ldots, K\}, K \in \mathbb{N}$, and randomly initialize an 'image' vector $\mathbf{w}_i \in E$, $i \in \{1, \ldots, N\}$ for each input vector $\mathbf{r}_i$. Typical choices for sampling distributions are: for structure-preserving visualization, use uniform distribution (e.g. in a 2D square, as used in fig. 1); for data clustering use several Gaussian distributions with different centers (e.g. located on the nodes of a regular simplex, as used in fig. 2.)

Step 3: Reconstruct the topology induced by the input data in O by moving the image vectors in the embedding space E using the computational scheme of a topology-preserving mapping T. The final positions of the image vectors $\mathbf{w}_i$ represent the output of the algorithm.

In the third step of the XOM algorithm, the topology-preserving mapping T can be considered as a free variable. A simple choice for T is Kohonen's self-organizing map algorithm [1], e.g. in its basic incremental version. Here, the image vectors $\mathbf{w}_i$ are incrementally updated by a sequential learning procedure. For this purpose, the neighborhood couplings between the input data items are represented by a so-called cooperativity function ψ. A typical choice for ψ is a Gaussian

$$\psi(\mathbf{r}, \mathbf{r}'(\mathbf{x}(t)), \sigma(t)) := \exp\left(-\frac{(\mathbf{r} - \mathbf{r}'(\mathbf{x}(t)))^2}{2\sigma(t)^2}\right). \tag{1}$$

In the XOM context, $\mathbf{r}'(\mathbf{x}(t))$ represents the 'best-match' input data vector. For a randomly selected sampling vector $\mathbf{x}(t) \in E$, this best-match input data vector is identified by $\|\mathbf{x} - \mathbf{w}_{\mathbf{r}'}\| = \min_{\mathbf{r}} \|\mathbf{x} - \mathbf{w}_{\mathbf{r}}\|$. Once the best-match input data vector $\mathbf{r}'(\mathbf{x}(t))$ has been identified, the image vectors $\mathbf{w}_{\mathbf{r}}$ are updated in a sequential adaptation step according to the learning rule

$$\mathbf{w}_{\mathbf{r}}(t+1) = \mathbf{w}_{\mathbf{r}}(t) + \epsilon(t)\, \psi(\mathbf{r}, \mathbf{r}'(\mathbf{x}(t)), \sigma(t))\, (\mathbf{x}(t) - \mathbf{w}_{\mathbf{r}}(t)), \tag{2}$$

where t represents the iteration step, $\epsilon(t)$ a learning parameter, and $\sigma(t)$ a measure for the width of the neighborhood taken into account by the cooperativity function ψ. In general, $\sigma(t)$ as well as $\epsilon(t)$ are changed in a systematic manner depending on the number of iterations t by some appropriate annealing scheme, e.g. an exponential decay with t, as used in the examples of this paper. The algorithm is terminated, once a problem-specific cost criterion is satisified, or a maximum number of iterations has been completed.

Although the above computational scheme formally resembles Kohonen's self-organizing map algorithm, there are deep differences between both approaches. Specifically, the meaning of the variables $\mathbf{r}$, $\mathbf{w}$, and $\mathbf{x}$ *completely differs* in the Exploration Machine: Whereas in Kohonen's algorithm the sampling vectors $\mathbf{x}$ represent the input data, this role is attributed to the vectors $\mathbf{r}$ in XOM. As an important consequence, in contrast to Kohonen's self-organizing map algorithm, each image vector $\mathbf{w}$ is attributed to its own specific input data vector. Hence, no implicit approximation of input data items by image vectors is involved. Instead, sampling and adaptation of the image vectors is entirely restricted to the embedding space. In other words, XOM completely inverts the role of input data and structure hypotheses, given the conventions of topology-preserving mappings as known from the literature. These differences induce that (i) the dynamics

ESANN'2009 proceedings, European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning. Bruges (Belgium), 22-24 April 2009, d-side publi., ISBN 2-930307-09-9.

of self-organization is formulated *directly* in the embedding space E in which structure formation occurs, and not indirectly via movements in the space O of the high-dimensional input data. This can lead to substantial computational savings, see below. Second (ii), the coupling of the movements of the image vectors is now governed by the actual distance topology of the input data and *not* by the possibly inaccurate structure hypothesis as in existing approaches[1]. As in topology-preserving mappings, we still need a structure hypothesis. But now it is succinctly spelled out in the choice of the sampling distribution and its underlying space E that govern the exploration movements of the image vectors *without* affecting their interactions.

The new scheme endows XOM with a surprising flexibility to contribute to different domains of scientific data analysis and visualization, such as clustering, see below. It also induces favorable algorithmic properties: In particular, the formulation of the dynamics in the embedding space entails a substantial reduction of computational complexity in comparison to the self-organizing map, as the best-match search in each iteration step does not require computational operations in the high-dimensional input data space, but now occurs in the usually low-dimensional embedding space. This leads to considerable savings in computation time when compared to the self-organizing map, specifically in the case of high-dimensional real-world data, such as for the embedding example of the whole-genome gene expression data in fig. 1.

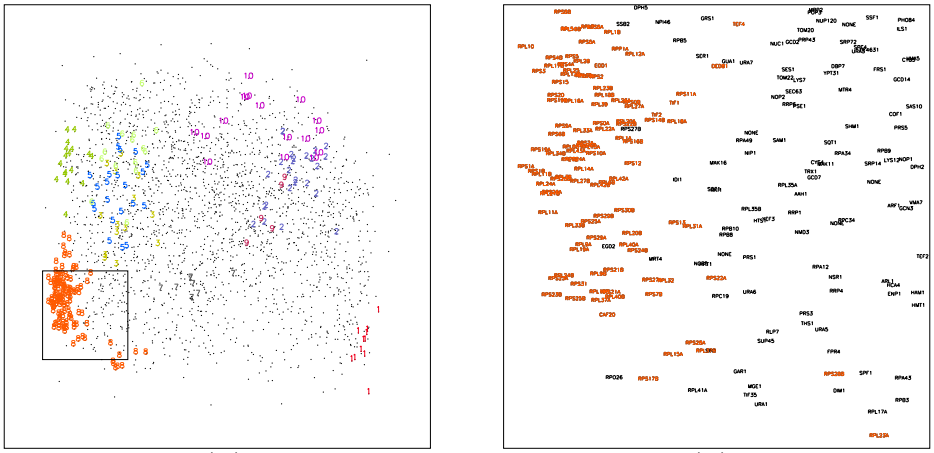
Although the Exploration Machine has originally been invented as a novel method for structure-preserving dimensionality reduction, it is essential to realize that it can be applied to other domains of data analysis as well. Data clustering, for example, can be performed by exploiting the flexibility to specify arbitrary structure hypotheses in step 2 of the algorithm. Here, the key idea is to simply select the sampling vectors from non-uniform distributions, e.g. from a mixture of several (e.g. Gaussian) distributions centered at different positions in the embedding space. After running the XOM algorithm, the image vectors can be assigned to these distributions, e.g. by computing and comparing the distances of the image vectors to the centers of the respective distributions. For instance, in the clustering example of section 3, fig. 2 below, we used a structure hypothesis of 36 univariate 35-dimensional Gaussian distributions located on the vertices of a regular simplex in $\mathbb{R}^{35}$.

For further analyses, extensions, and variants of the Exploration Machine, such as related to computational complexity, convergence properties, free parameter selection, supervised learning, analysis of non-metric data, out-of-sample extension, and constrained incremental learning, we refer to [2].

3 Experiments

Visualization of Genome-Wide Expression Patterns: We used the Exploration Machine to visualize genome-wide expression patterns by structure-

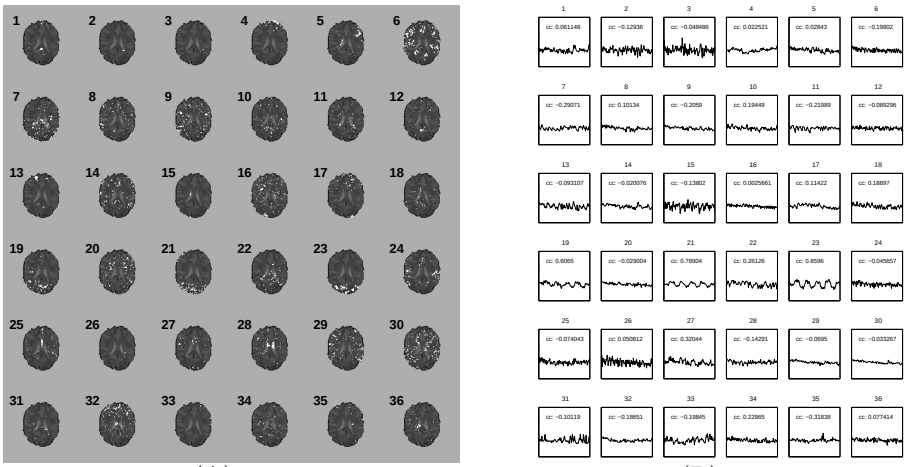
[1] A typical choice for defining such structure hypotheses in topology-preserving mappings is to use two-dimensional, discrete, periodic (e.g. quadratic or hexagonal) grids.



(A) (B)

Fig. 1: Visualization of genome-wide expression profiles by the Exploration Machine. (A) Genome map created by nonlinear XOM embedding of 2467 79-dimensional gene expression profiles in the yeast *Saccharomyces cerevisiae* obtained from DNA microarray hybridization experiments, data published in [3]. (B) Enlarged section of the genome map depicted by the inset in (A).

preserving dimensionality reduction. Fig. 1 (A) shows a genome map created by nonlinear XOM embedding of gene expression profiles in the yeast *Saccharomyces cerevisiae* obtained from image data of DNA microarray hybridization experiments. The data is taken from [3], where it is described in detail. It includes 2467 79-dimensional vectors representing concatenated time courses obtained for all genes functionally annotated in the *Saccharomyces* Genome Database [4]. The expression profiles were average-corrected and scaled to unit variance. – Each of the 2467 points on the map represents the 79-dimensional expression profile of a single gene. Specific groups of genes related to each other with respect to their biological function according to a cluster annotation by Eisen et al. [3] are color-coded and labeled by numbers. In detail: '1': spindle pole body assembly and function, '2': the proteasome, '3': mRNA splicing, '4': glycolysis, '5': the mitochondrial ribosome, '6': ATP synthesis, '7': chromatin structure, '8': the ribosome and translation, '9': DNA replication, and '10': the tricarboxylic acid cycle and respiration. Fig. 1 (B) shows the enlarged section of the genome map depicted by the inset in (A). For the practical use of XOM genome maps as presented in Fig. 1, appropriate graphical user interfaces can easily be implemented that supply additional annotation information on the map when needed. Correspondingly, plots of the individual expression profiles can easily be projected onto the map. – A notable result from the visualization in fig. 1 is that genes of similar biological function are collocated on the map. We quantitatively compared our result of fig. 1 with the results obtained by several other embedding algorithms using Sammon's error function [5] as a criterion for em-



(A) (B)

Fig. 2: Exploration Machine cluster analysis in human brain mapping, based on a functional MRI visual stimulation experiment. (A) Cluster assignment maps. White regions denote pixels assigned to a respective cluster. (B) Cluster-specific prototypical time-series, i.e. 'codebook vectors' and cluster-specific correlation coefficients between codebook vectors and stimulus function. Cluster numbers correspond to assignment maps in (A).

bedding quality: We obtained error values of $5.91 \cdot 10^5$ (1.00) for XOM, $6.50 \cdot 10^5$ (1.10) for Sammon's mapping, $6.56 \cdot 10^5$ (1.11) for Principal Component Analysis (PCA), and $7.24 \cdot 10^5$ (1.22) for SOM, where numbers in brackets indicate relative values compared to the XOM result. Computation times were 72 s for XOM, 216 s for Sammon's mapping, 2 s for PCA, and 881 s for SOM. Our results indicate XOM's applicability to provide fast and concise structure-preserving visualization of large biomedical data collections, as exemplified by whole-genome microarray data.

Functional MRI for Human Brain Mapping: In the previous section, the Exploration Machine has been successfully applied to structure-preserving *dimensionality reduction.* To demonstrate its applicability to *data clustering* as well, we performed exploratory functional MRI analysis for human brain mapping in a visual stimulation experiment. Here, the basic idea is to group pixels according to their similarity of pixel-specific signal dynamics time-series. Experimental protocols for image acquisition of this data set have been published in [6]. Each functional MRI slice includes approximately $5-10 \cdot 10^3$ pixels, with a number of 98 acquisitions over a time of 300 s. Thus, the task is to cluster several thousand time-series vectors in $\mathbb{R}^{98}$. Figures 2 (A) and (B) show an example of cluster assignment maps and corresponding cluster-specific prototypical signal-time series, so-called 'codebook vectors', that can be interpreted as the average time-series of all the pixels belonging to a specific cluster. As can be seen from the figures, clusters 21 and 23 clearly identify task-related activity in the

visual cortex, reflected by the high correlation between codebook vectors and the box-car shaped stimulus function used in this experiment. Cluster 24 includes pixels representing cerebrospinal fluid of internal ventricles, whereas cluster 4 is indicative for a through-plane motion artifact. A quantitative ROC analysis revealed areas under ROC curves of 0.984 ± 0.03 for XOM, 0.983 ± 0.02 for Minimal-Free-Energy VQ [6], and 0.979 ± 0.02 for SOM for the detection of task-related activation. We conclude that our method is well-suited to perform high-dimensional cluster analysis of functional MRI data yielding competitive results comparable to those obtained by established methods, e.g. [6].

4 Conclusion

In this paper, we have introduced the Exploration Machine as a novel learning approach which can be applied to the analysis of multidimensional data. We have shown that XOM is capable of visualizing whole-genome microarray gene expression data and that it can be used for clustering of functional MR image time-series in human brain mapping. For both applications, we found that XOM yields competitive results when compared to established methods known from the literature. It can be shown that XOM introduces a generalization of various previous learning approaches, such as [7], [8], i.e. XOM includes these as special cases within a general framework that proposes *to invert topology-preserving mappings as a fundamental pattern recognition approach.* As XOM can simultaneously contribute to different domains of machine learning, namely both dimensionality reduction *and* clustering, it may serve as a useful novel versatile method for exploratory data analysis throughout science and engineering.

References

[1] Kohonen, T., [*Self-Organizing Maps*], Springer, Berlin, Heidelberg, New York, 3rd ed. (2001).

[2] Wismüller, A., *Exploratory Morphogenesis (XOM): A Novel Computational Framework for Self-Organization*, Ph.D. thesis, Technical University of Munich, Department of Electrical and Computer Engineering (2006).

[3] Eisen, M., Spellman, P., Brown, P., and Botstein, D., "Cluster analysis and display of genome-wide expression patterns," *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868 (1998).

[4] Cherry, J., Ball, C., Weng, S., Juvik, G., Schmidt, R., Adler, C., Dunn, B., Dwight, S., Riles, L., and Mortimer, R. *Nature* **387**, 67–73 (1997).

[5] Sammon, J., "A nonlinear mapping for data structure analysis," *IEEE Transactions on Computers* **C 18**, 401–409 (1969).

[6] Wismüller, A., Lange, O., Dersch, D., Leinsinger, G., Hahn, K., Pütz, B., and Auer, D., "Cluster analysis of biomedical image time-series," *International Journal of Computer Vision* **46**(2), 103–128 (2002).

[7] Lee, J. A. and Verleysen, M., "Nonlinear projection with the Isotop method," in [*Proc. ICANN, LNCS 2415*], Dorronsoro, J., ed., 933–938, Springer, Berlin (2002).

[8] Wismüller, A., Vietze, F., Dersch, D., Hahn, K., and Ritter, H., "The deformable feature map — adaptive plasticity in function approximation," in [*Proc. ICANN'98*], Niklasson, L., Bodèn, M., and Ziemke, T., eds., **1**, 222–227, Springer-Verlag, London, Berlin, New York (1998).