Chapter 18

Other projects

18.1 Mixture density from autonomous experts

Jarkko Salojärvi and Samuel Kaski

In mixture density modeling it is assumed that each data sample is generated by one of a set of independently operating mixture components, the whole data set being a mixture of samples from them. When fitting a mixture of models or "experts" to data, the modeling task is broken into partially overlapping subtasks assigned to relatively simple experts. This increases analytical tractability and makes the results more easily understandable, a topic especially important in tasks where the goal is to summarize large data sets and gain understanding of their essential characteristics.

Although each expert generates data independently, an expert within a mixture cannot be fitted autonomously to data because the contributions of all the other experts must be known. In [1,2] we introduce a variant of the mixture model in which the experts can learn autonomously by maintaining simplified, periodically updated representations of the other experts. If the representations are simpler than the original experts themselves, the complexity of each expert will be considerably lower than that of the whole mixture. By assigning the autonomously operating experts to different processors, computational load can be distributed. The autonomous learning procedure can be interpreted so that each expert modulates its learning by its estimate of the probability that it is responsible for the data.

The principle is demonstrated with toy data in Figure 18.1. We have applied the autonomous learning to several data sets [2].

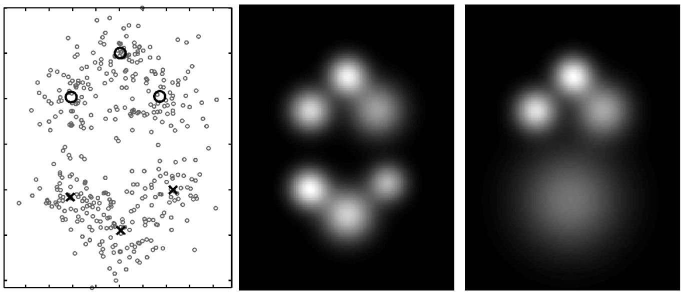

Figure 18.1: A toy example demonstrating the autonomous experts. Left: The data (small circles) was modeled with two autonomous experts, each consisting of three Gaussians. The means of the Gaussians are shown with circles for expert 1 and crosses for expert 2. Center: the density generated by the mixture. Right: the density plot shows how the first expert sees the data. The large Gaussian at the bottom represents the other expert.

References

[1] S. Kaski and J. Salojärvi. Generative mixture modeling by autonomous estimators. In N. Baba, L. C. Jain, and R. J. Howlett, editors, *Knowledge-Based Intelligent Information Engineering Systems & Allied Technologies, Proceedings of KES'2001*, volume 1, pages 250–254. IOS Press, Amsterdam, 2001.

[2] J. Salojärvi and S. Kaski. Mixture density from autonomous experts. *International Journal of Knowledge Based Intelligent Engineering Systems*, 6(1):48–55, January 2002.

18.2 Self-organizing map-based information visualization is trustworthy

Jarkko Venna and Samuel Kaski

One of the main uses of Self-Organizing Maps is information visualization in exploratory data analysis. A Self-Organizing Map of a large unknown multivariate data set gives a first comprehensive overview of the similarity relationships of the data samples and cluster structures in the set.

Several measures have been proposed for comparing nonlinear projection methods but so far no comparisons have taken into account one of their most important properties, the trustworthiness of the resulting neighborhood or proximity relationships. In visualizations it is crucial that the visualized proximities can be trusted upon: If two data samples are close to each other on the display they should be close-by in the original space as well. We have proposed to measure trustworthiness by the number of samples arriving from outside of the original neighborhood to a small neighborhood on the display (more accurately: the rank distance from the neighborhood [1]).

A sample experiment is shown in Figure 18.2**a**. For the important small neighborhoods (small k), neighborhood relationships visualized by the Self-Organizing Map (SOM) and its variant, the Generative Topographic Mapping (GTM), are more trustworthy than visualizations produced by traditional multidimensional scaling-based nonlinear projection methods. Yet, Self-Organizing Maps preserve the original neighborhoods with a comparative accuracy (Fig. 18.2**b**).

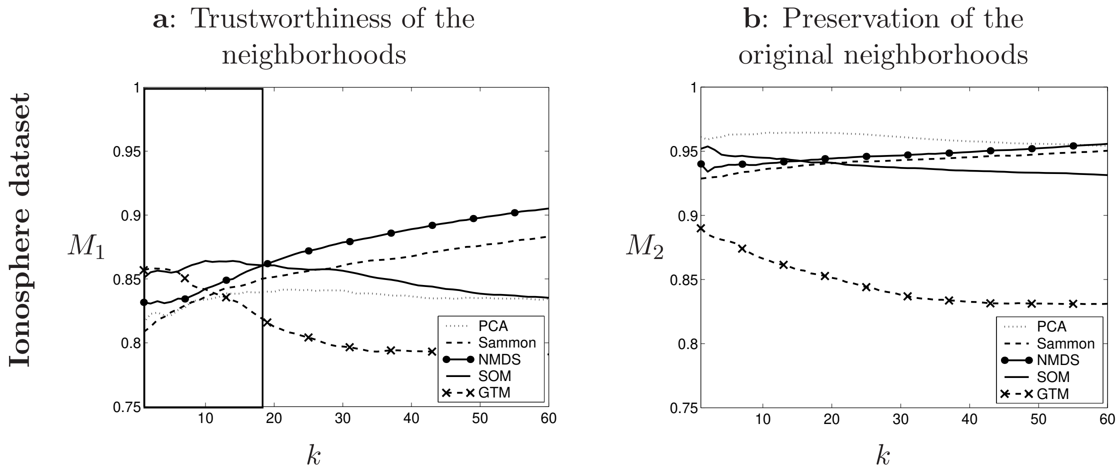

Figure 18.2: Trustworthiness (M_1) of the neighborhoods after projection (**a**) and preservation (M_2) of the original neighborhoods (**b**) for a sample data set as a function of the neighborhood size k. The SOM outperforms the MDS-based methods in the boxed region. PCA: principal component analysis, Sammon: Sammon's mapping, NMDS: non-metric multidimensional scaling.

References

[1] J. Venna and S. Kaski. Neighborhood preservation in nonlinear projection methods: An experimental study. In G. Dorffner, H. Bischof, and K. Hornik, editors, *Artificial Neural Networks—ICANN 2001*, pages 485–491. Springer, Berlin, 2001.

18.3 SOM in detecting states of sleep and wakefulness in polysomnographic data of mentally retarded persons

Sirkka-Liisa Joutsiniemi, Maija-Liisa Laakso, Lea Leinonen, Jussi Nurminen, Teuvo Kohonen

Interview and observation based studies show that sleep problems are common in the mentally retarded [1], [3]. Polysomnography is 'the golden standard' for all objective methods in sleep research. There are often difficulties, however, in applying this method in the investigations on subjects with damaged brains. The features used in the classification of wake and sleep stages of healthy subjects may be lacking or deformed, or there may be abnormalities disturbing the definitions. The patient group is heterogeneous and no common rules apply to them.

In collaboration with Rinnekoti Foundation, we are developing a SOM-based method to help the clinicians in the analysis of the polysomnography of the mentally retarded. The ability of the SOM to learn unsupervised the regularities in the individual data and visualize the structure of it offers a good basis for such a tool [2] (Fig. 18.3). A prerequisite for efficient SOM are new signal features suitable for detecting wake and sleep states in the mentally retarded. The work concerning the feature extraction is going on.

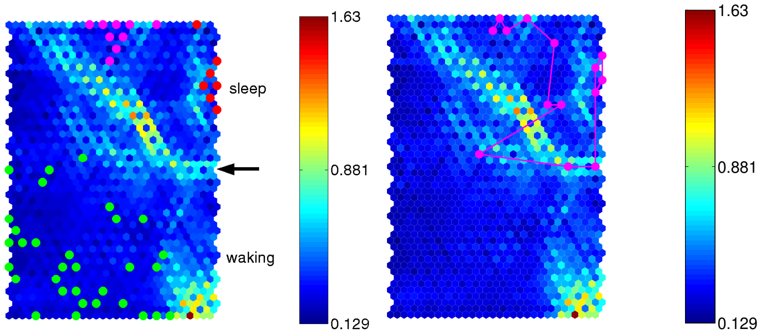

Figure 18.3: The SOM diagrams of the 24-h polygraphy of a healthy subject. *Left:* The border between waking and sleep areas is indicated by an arrow. On the sleep area, the red dots on the right are projections of 60 s deep sleep (S4) episodes of the polygraphy, dots on the left of the same area are projections of REM sleep episodes. The lower part of the map receives projections of waking (green dots). *Right:* The trajectory of the moving projection of transition from S4 to REM sleep.

References

[1] Lindblom N, Heiskala H, Kaski M, Leinonen L, Nevanlinna A, Iivanainen M, Laakso M-L: Neurological impairments and sleep-wake behaviour among the mentally retarded. *Journal of Sleep Research 10, 309-318, 2001.*

[2] S.-L. Joutsiniemi, J. Nurminen, T. Kohonen SOM in detecting states of sleep and wakefulness in polysomnographic data. *Proceedings of the Fourth International Conference on NNESMED 2001*, pages 97–101.

[3] Lindblom N, Heiskala H, Kaski M, Leinonen L, Laakso M-L: Sleep fragmentation in mentally retarded people decreases with increasing daylength in spring. *Chronobiology International 19, 1-19, 2002 (in press).*

18.4 Independent variable group analysis

Krista Lagus, Esa Alhoniemi, Harri Valpola

The goal of unsupervised learning is to extract an efficient representation of the statistical structure implicit in the observations. A good model is both accurate and simple in terms of model complexity, i.e., it forms a *compact representation* of the input data.

In problems with a large number of diverse observations there are often groups of input variables that have strong mutual dependences within the group but which can be considered practically independent of the input variables in other groups. It can be expected that the larger the problem domain, the more independent groups there are. Estimating a model for each independent group separately produces a more compact representation than applying the model to the whole set of variables. Compact representations are computationally beneficial and, moreover, offer better generalization.

We present an approach called independent variable group analysis (IVGA), where the dependences of variables within a group are modeled, whereas the dependences between the groups are neglected [2]. Usually such variable grouping is performed by a domain expert, prior to modeling with automatic, adaptive methods. As expert knowledge may be unavailable, or expensive and time-consuming, automating the task can considerably save resources. The IVGA is a practical, efficient and general approach for obtaining compact representations that can be regarded as sparse codes, as well.

The $\mathrm{IVGA}_{\mathrm{VQ}}$ algorithm

Any IVGA algorithm consists of two parts, (1) grouping of variables, and (2) construction of a separate model for each variable group. An independent variable grouping is obtained by comparing models with different groupings using a suitable cost function. In principle any model can be used, if the necessary cost function is derived for the model family.

We have applied the IVGA approach for a situation where the dependences within variable groups are modeled using vector quantization (VQ), and derived the necessary cost function for model optimization with VQ. For optimization of the model we use the variational EM-algorithm (cf. e.g. [1,3]).

Since it is generally computationally prohibitive to try all the different groupings of D variables into any number of groups, some heuristic optimization strategy has to be utilized. We have used the following strategy:

- **Initialize:** Assign each variable to its own group. Model each group using VQ and calculate total model cost.
- **Repeat until time limit or model cost limit:**
  - Consider a change in the grouping of variables:
    - Move a variable from group to another
    - merge two groups
    - Run $\mathrm{IVGA}_{\mathrm{VQ}}$ recursively for a variable group or the union of two groups (split or merge+split)
  - Model each group using a VQ and calculate total model cost. If cost improved, accept the change.

As a cost function one can use negative log-likelihood of the data given the model, namely $-\ln p(\mathbf{x}|H)$. The total model cost L_{tot} needed for comparing variable groupings is the sum of costs of individual variable groups $L_{\rm tot} = \sum_g L_g = \sum_g -\ln p(\mathbf{x}_g|H_g)$, where g is the index of a group of variables, and x_g and H_g are the data and the model related to that variable group, respectively. Regarding the VQ model and the variational EM algorithm used for optimization of the model parameters the interested reader is referred to [2].

Experiments

Experiments were carried out to (1) verify the general IVGA principle by comparing the obtained groupings to a known categorization of variables, and (2) to study the performance of the presented algorithm by comparing model costs obtained using IVGA_{VQ} and regular VQ.

The data set consisted of 1000 images from the PicSOM project (see Chapter 7). Three kinds of features (variables) had been calculated to represent each image: FFG, RGB, and texture. It is reasonable to assume that dependences between variables in different categories are weak (cf. e.g. [4]). The original images were represented by 144 features, of which 50 were randomly chosen here: 27 FFT features (feature set A), 7 RGB features (feature set B), and 16 texture features (feature set C).

A summary of the results is shown in Table 18.1. The VQ(A+B+C) denotes the running of the regular VQ for the combined set of features A, B, and C. In particular, the results show that IVGA_{VQ}(A+B+C) obtained a clearly better model than plain VQ(A+B+C).

Table 18.1: Summary of the results of the experiments. The groupings obtained using IVGA_{VQ} contain variables from one category only.

Experiment	**Total cost**	**# VQs**	**# Params**
1. VQ(A+B+C)	-138115.5	1	2200
2. VQ(A) + VQ(B) + VQ(C)	-145796.2	3	1045
3. IVGA_{VQ}(A+B+C)	-147206.8	12	618
4. IVGA_{VQ}(A) + IVGA_{VQ}(B) + IVGA_{VQ}(C)	-147934.3	9	712

In conclusion, the experimental results show that it is worthwhile to group variables according to independence, and that the presented algorithm is able to do this and in doing so, obtains more compact models.

References

[1] G. E. Hinton and D. van Camp. Keeping neural networks simple by minimizing the description length of the weights. In *Proceedings of the COLT'93*, pp. 5–13, Santa Cruz, California, USA, July 26–28, 1993.

[2] K. Lagus, E. Alhoniemi, and H. Valpola. Independent Variable Group Analysis. In *Proceedings of International Conference on Artificial Neural Networks - ICANN 2001*, pp. 203–210. Vienna, Austria. Springer, 2001.

[3] H. Lappalainen and J. W. Miskin. Ensemble learning. In M. Girolami, ed., *Advances in Independent Component Analysis*, pp. 76–92. Springer-Verlag, Berlin, 2000.

[4] E. Oja, J. Laaksonen, M. Koskela, and S. Brandt. Self-organizing maps for content-based image database retrieval. In E. Oja and S. Kaski, eds., *Kohonen Maps*, pp. 349–362. Elsevier, 1999.

18.5 Modeling visual attention

Teuvo Kohonen

Introduction

The term *attention* has been used in connection with many different psychological, behavioral, and physiological states and actions. Here we may simply define attention as a set of those cognitive functions by which primary sensory occurrences are selected and emphasized for perception.

Many phenomena related to attention have been described already by the ancient and medieval philosophers such as Aristotle, Lucretius, and Descartes. Into the scientific disciplines that preceded the modern psychology, attention was introduced in the 1740s by Wolff [1], and one of his observations was that the greater the attention, the smaller part of the visual field to which it extends.

Over the years, also many different terminologies and descriptions have been used [2], but at the moment there seems to be general agreement on at least the following fundamental features: 1. Attention is usually triggered by the onset of some external alerting occurrence (although it can also be evoked by various cues in mental tasks), 2. After the initial phase, attention is usually actively directed and narrowed, 3. Attention may be shifted involuntarily (e.g. by distractors), 4. The maintenance of attention for extended periods is difficult and easily interfered, 5. Cognitive interpretation of a prior stimulus has a biasing or *priming* effect on the direction of attention.

The objective of the work in presentation is very concrete: to give, by means of very simple modeling approaches, an explicit explanation to some of the most salient phenomena especially in visual attention, namely: 1. Activation of a part of the visual field during the fixation of the gaze, whereupon the width of the activated part will only depend on the "information content" of that part at a given relative resolution, 2. Narrowing of the activated field at more concentrated watching, whereupon "concentration" only means small eye movements, voluntary or involuntary.

In other words, I believe that at least the above two phenomena may have a very simple mechanistic interpretation, without any "consciousness" or "will" yet involved with them. Contrary to that, acts like involuntary shifting, priming, and even coordination of the muscular actions during attention may involve integrated cognitive and sensorimotor mechanisms and fall outside the scope of this discussion.

Modeling assumptions

I shall start with a few basic assumptions for modeling. These assumptions about the (mammalian) visual system will first be expressed in a general but yet concrete form in order to allow for various explicit implementations.

(i) The visual signals are propagated from the retina to the visual cortex via the thalamus along pathways that preserve their topological order all the way through.

(ii) The density of the retinal ganglia varies within wide limits over the retina. The signal pathways that start in the center of the retina are able to carry signals that represent denser variations in the spatial domain, while the peripheral pathways are better fit to smoother variations in the spatial domain.

(iii) With the main signal pathways there are associated other functions that modulate the transmittance of the former to the signals. These functions consist of neural circuits, each one of them monitoring and controlling its own subset of signal pathways up to a certain radius in the lateral direction. The associated circuits analyze the local "information content" of the signals, e.g., both spatial and temporal variations of the signals in their neighborhood, at different resolution. The control circuits also *compete* mutually in the sense that preference for the transmission of signals is given to those pathways, the associated control circuit of which detects the greatest variations, in relation to its own resolution, in the signals in its neighborhood. The rest of the neighboring pathways is inhibited or somehow neglected.

(iv) If the gaze is directed at a small object or densely located details concentrated on a small area, and there are no large optic structures in its surround, the control circuits associated with the narrow central areas of the retina are activated stronger than the peripheral control circuits. Accordingly, the central pathways will then be activated for signal transmission. Conversely, the broader or smoother the optic structures of the image around the gaze, the broader signal pathways are activated.

(v) As the control circuits are assumed sensitive to temporal changes in the spatial signal patterns, too, then, if the image is translated by a small amount on the retina, the relative changes caused by this translation are much bigger in the control circuits that modulate signals in the central area, where the resolution is high, than in the large peripheral areas with low resolution, respectively.

(vi) Thus, when the gaze is shifted by a small amount, activation of the central pathways in relation to the peripheral ones is increased, which is experienced as narrowing of the visual field and concentration of attention.

Simulations

In computer simulations I have used photographic images. The control circuits were assumed to lie symmetrically around the assumed direction of the gaze. The image intensities of these parts were modulated by the control circuits.

In a more extensive simulation one might want to place control circuits and pathways all over the surface of the simulated "retina" and also let them overlap. In this simulation, for simplicity, I have only used four overlapping and competing control circuits that lay concentrically around the direction of the gaze, with their widths relating approximately as 1:2:4:8.

Every control circuit shall in the first place detect whether there occur local spatial variations in the image at low resolution, and the control circuit shall be selective to optic structures the dimensions of which are of the same order of magnitude as the diameter of the circuit itself. Such an analysis can be performed in several ways. One possibility is that each control circuit operates like a low-level feature filter that may be implemented by laterally interconnected interneurons. In the present simulations, a functionally defined, very effective and computationally light method was used. A small number of local averages of the signals over each hypothetical control circuit were computed, and then the *variance* of these local averages was evaluated. The pathway over which the (relative) variance was biggest was activated for transmission.

In order to simulate the "narrowing" of the visual field under concentrated attention, the control circuits were also made sensitive to temporal variations in the signals. In the simplest way this was done in simulation by subtracting from each local average a fraction of the average computed previously, and evaluating the variances from these temporal differences of the spatial patterns.

(a)

(b) (c)

Figure 18.4: (a) The original image. (b) A part of the image was selected and activated by the "winning" control circuit. (c) When the gaze was moved by a small amount towards a detail (pedestal of the statue), the activated part of the visual field was further narrowed.

References

[1] Wolff, C. *Psychologia rationalis.* (Officina Libraria Rengeriana, Frankfurt & Leipzig 1740).

[2] Wright, R.D. (ed.) *Visual attention* (Oxford University Press, New York & Oxford 1998).