Self-Organizing Map in Data-Analysis - Notes on Overfitting and Overinterpretation

Jouko Lampinen and Timo Kostiainen
Laboratory of Computational Engineering
Helsinki University of Technology
P.O.Box 9400, FIN-02015 ESPOO, FINLAND

Abstract

The Self-Organizing Map, SOM, is a widely used tool in exploratory data analysis. Visual inspection of the SOM can be used to list potential dependencies between variables, that are then validated with more principled statistical methods. In this paper we discuss the use of the SOM in searching for dependencies in the data. We point out that simple use of the SOM may lead to excessive number of false hypotheses. We formulate the exact probability density model for which the SOM training gives the Maximum Likelihood estimate and show how the model parameters (neighborhood and kernel width) can be chosen to avoid overfitting. The conditional distributions from the true density model offer a consistent way to quantify and test the dependencies between variables.

1   Introduction

The Self-Organizing Map, SOM, [3] has established its position as a widely applied tool in data-analysis and visualization of high-dimensional data. Within other statistical methods the SOM has no close counterpart, and thus it provides a complementary view to the data. The SOM carries some similarities to multidimensional scaling methods (such as the Sammon's method), principal curve projection and other non-linear latent variable methods. For a recent review see e.g. [2]. Compared to the SOM, the other methods are typically better understood theoretically, as they are based on optimizing some well defined objective functions or model fitting criteria, while the SOM is defined constructively by the training algorithm and has its roots in a simplified model for the self-organization process in biological neural networks. The SOM is, however, the most widely used method in this category, because it provides some notable advantages over the alternatives. These include, e.g., ease of use, especially for inexperienced users, and very intuitive display of the data projected on to a regular two-dimensional slab, as on a sheet of a paper.

The main potential of the SOM is in *exploratory data analysis*, which differs from standard statistical data analysis in that there are no presumed set

of hypotheses that are validated in the analysis. Instead, the hypotheses are generated from the data in the data-driven *explor atory* phase and validated in the *confirmatory* phase.

There are some problems where the exploratory phase may be sufficient alone, such as visualization of data without more quantitativ e statistical inference upon it. In practical data analysis problems the most common task is to search for dependencies between variables. In such a problem, the SOM can be used for getting insight to the data and for the initial search of potential dependencies. In general the findings need to be validated with more classical methods, in order to assess the confidence of the conclusions and to reject those that are not statistically significant.

In this contribution we discuss the use of the SOM in searching for dependencies in the data. First we formulate the probability density model for which the SOM training gives the Maximum Likelihood estimate, and give a method for ev aluating the normalized probability values from the model. This allows using standard methods like cross-v alidation to select the model complexity (neighborhood and kernel widths) so that the model is not overfitted to the data sample.

Then w e show how conditional probabilities from the model can be computed to estimate the dependencies between variables. We point out that visual inspection of the SOM may lead to excessive number of false hypotheses, as information about independence of the variables may be distributed in a w ay that is not easily visible on the map.

In many alternatives for the SOM these issues are much easier to cope with, especially when the discontinuous Voronoi regions are replaced by regular continuous kernels, such as in the Generative Topographic Mapping [1], and in the Bay esian SOM [5], where the neighborhood is replaced by Gaussian smoothing kernel on the unit positions.

2   Probability Density Estimate by the SOM

In this section we deriv e the probability density model, for which the minimization of the SOM cost function giv es maximum likelihood estimate. The cost function, that is stochastically minimized by the SOM, is [4]

$$E = \sum_{n=1}^{N} \sum_{r=1}^{M} H(b(x) - r) \|x^n - m^r\|^2, \tag{1}$$

where $X = \{x^n\}, n = 1, \ldots, N$ is the discrete data sample, r is the index (or position) of a unit in the SOM, $b(x)$ is the index of the best matching unit for x, m^r is the reference vector of the unit r, and $H()$ is the neighborhood function.

The Maximum Likelihood (ML) estimate is based on maximizing the probability of the data given the model, which is equal to minimization of the negative log-likelihood cost function. Let $p(X|m, H)$ be the likelihood of the data X from the SOM model with codebook m and neighborhood H. The negative log-likelihood is $L = -\log p(X|m, H)$ and by setting it proportional to the cost

function in Eq.1 yields

$$p(X|m, H) = Z' \exp(-\beta E) = Z' \exp(-\beta \sum_n \sum_r H(b(x) - r)\|x^n - m^r\|^2). \quad (2)$$

Here we have replaced the normalizing constant Z' and the width of the kernels β that are not needed in the ML estimate of the codebook m, but are necessary for the complete density model. Let us define the basic SOM density kernel as

$$\varphi(x, r, H) = \exp\left(-\beta H(b(x) - r)\|x^n - m^r\|^2\right). \quad (3)$$

It is a Gaussian kernel centered at unit r with standard deviation $(2\beta H(b(x) - r))^{-1/2}$, which is larger for the Voronoi cells far away on the map, as illustrated in Fig. 1. The density of x can now be written as

$$p(x|m, H) = Z \prod_r \varphi(x, r, H) \quad (4)$$

and the likelihood in Eq. 2 is $p(X|m, H) = Z^N \prod_n \prod_r \varphi(x^n, r, H)$.
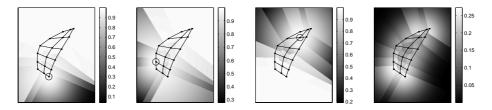

Figure 1: Example of density estimate due to the SOM. Gaussian neighborhood width and β were chosen to give best fit to validation data. The plots show three different kernels $\varphi(x, r, H)$, centered on the map units that are marked by circles. The last plot shows the resulting density estimate, which is the product of all the 18 different kernels.

To compare the likelihood of models with different neighborhoods the normalizing constant Z needs to be evaluated from equation

$$\int p(x|m, H)dx = Z \sum_r \int_{x \in V_r} \prod_s \varphi(x, s, H)dx = 1, \quad (5)$$

where the integration over x is decomposed to sum of integrals over all the Voronoi cells. Due to the discontinuity of the kernels in the Voronoi cell boundaries the integrals cannot be computed in closed form, but they can easily be approximated numerically with rejection sampling Monte Carlo method as follows. 1) For each cell r, draw L samples from proposition distribution $Q(x) = \varphi(x, r, H)$. 2) Accept a sample x if it is inside the Voronoi cell V_r with probability $\prod_s \varphi(x, s, H)/Q(x)$. 3) The volume $\int_{x \in V_r} \prod_s \varphi(x, s, H)dx$ is the fraction $L_{accepted}/L$ of the volume of Q, which is $(\pi/\beta)^{d/2}$ where d is the dimension of the data space. With the normalized likelihood one can choose the neighborhood and β to maximize the likelihood of the training data, or preferably, if the model has large number of free parameters, to maximize the likelihood of validation data. Fig. 2 shows the SOM density models for various neighborhood sizes, with ML estimate for β.
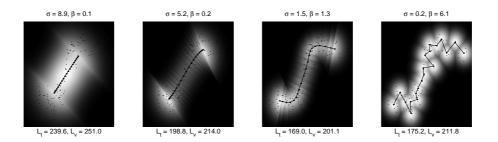
Figure 2: SOM density models for different widths σ of the Gaussian neighborhood. From the total likelihood of validation data the optimal neighborhood can be chosen to avoid o verfitting.The negative log-lik elihoods are denoted by L_t for training data and L_v for validation data.

3 SOM and Dependence Between Variables

In practical data analysis problems the most common task is to search for dependencies betw een v ariablesStatistical dependence means that the value of a variable can be predicted to a certain accuracy from one or more of the other variables, or in other words, the conditional distribofion the v ariable is dependent on the values of the other variables.

The most simple goal is to look for pairwise dependencies, where a variable depends only on one other variable. For such a problem the advan tage of the SOM is rather marginal, as simple correlation analysis is sufficient for linear case and in non-linear case direct viewing or modelling of the pairwise data samples is no more laborious than pairwise visual inspection of the component planes.

The tough problem is to search for non-linear dependencies between multiple variables. F rom the SOM these are searc hed b y looking for regions on the map with correlating v alues for assumed dependent variables. Such a region is interpreted as a hypothesis that the v ariables are dependent giv en that the other variables are close to the corresponding values in the reference vectors.

It is v ery importan t to notice that any conclusions drawn from models overfitted to the data sample are not guaranteed to generalize to any other situation. Thus analysis of statistical dependencies requires some way, heuristic or more disciplined, to avoid overfitting, as in all statistical modelling.

Given that the model is not badly overfitted, the problem in visual inspection of the SOM is that in general *the dependencies betwe en variables an not be seen from local analysis of the SOM*. That is, even if v ariables, say, x_1 and x_2 both ha velarge values at map unit r, that alone does not sho wthat the variables ha ve an y mutual dependence. As a simple example, consider tw o-dimensional uniform distribution $x_1, x_2 \sim U(-1, 1)$. A 2×2 SOM with zero neighborhood w ould ha ve componen t planes (in any order of the columns and rows)

$$M_1 = \begin{bmatrix} -0.5 & -0.5 \\ 0.5 & 0.5 \end{bmatrix} \qquad M_2 = \begin{bmatrix} -0.5 & 0.5 \\ -0.5 & 0.5 \end{bmatrix} \qquad (6)$$

The coincidence of high values in unit M^{22} and low values in M^{11} are only result of the vector quantization. T o see that high values in M^{22} do not indicate

dependence between x_1 and x_2, one must observe that high value for x_1 occurs also in M^{12} with low value for x_2 (i.e., tallied over the map, high value for x_1 indicates either high or low value for x_2).

In high dimensional space the visual inspection of the dependencies becomes more difficult, when the map folds and the variable ranges are distributed around the map, as illustrated in the Figs.3 and 4.
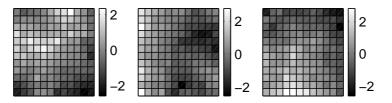

Figure 3: Example of SOM trained on purely random data. The independence of the variables in the component level display is not trivial to observe. One might, for example, erroneously conclude that high values of x_3 would indicate low values of x_2. Here the neighborhood is trained down to zero.
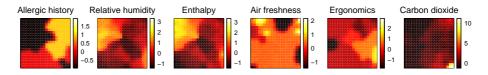

Figure 4: Example of real data analysis. In the case study ,the dependence of *Air freshness* of the other variables was investigated. In the final analysis all hypotheses were rejected using methods like RBF models, Bayesian neural networks, etc.

Quantitative analysis can be carried out by computing the conditional distributions from the density model. From the joint probability in Eq.4 conditional distribution for variable x_j can be numerically computed as $p(x_j|x_{\backslash j}, m, H) = p(x|m, H)/ \int p(x|m, H)dx_j$, where $x_{\backslash j}$ denotes the vector x with element j excluded. The prediction of x_j can similarly be computed as the conditional mean $E[x_j|x_{\backslash j}, m, H]$.

Reducing the kernel width to zero, $\beta \to \infty$, gives an important special case. The conditional density is then sharply peaked at the value of the "outputs" x_j in the best matching unit for the "inputs" $x_{\backslash j}$. The posterior mean $E[x_j|x_{\backslash j}]$ gives the same value as the nearest neighbor regression with the SOM reference vectors, producing piecewise constant estimate with large variance.

In Fig. 5 the 3D random data in Fig. 3 is analyzed by 6×6 SOM. The optimal width of the Gaussian neighborhood function is $\sigma = 4.2$, which is rather large, suggesting independent variables. From the figures it is clear that true conditional density is better estimate than the nearest neighbor rule and the model with correct complexity gives much better results. With visual inspection of the map it is difficult to perceive the mean or shape of the conditional distributions and thus the reliability of the conclusions is rather impossible to assess.
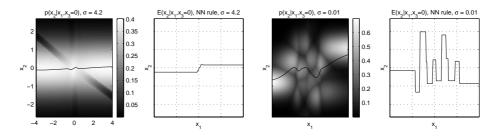
Figure 5: Conditional densities from a SOM trained on random independent data. The left figures show the conditional density and the nearest neighbor prediction for optimal neighborhood $\sigma = 4.2$, and the right figures for small neighborhood $\sigma = 0.01$, respectively.

4   Conclusions

We have discussed some difficulties in using SOM for analysing the statistical dependencies between variables. Even in lack of any dependence, high or low values of the variables on the map coincide, so that visual inspection easily suggests dependence. Thus it is necessary to validate the hypotheses using other statistical methods.

We derived the probability density model behind the SOM, and showed how the correct model complexity can be determined and the potential dependencies analyzed. Analysis of the model requires tedious numerical integrations, so that other more direct models may be more appropriate for the task.

Ultimately, to produce a SOM that is optimally suited for visual inspection, we conclude that the model parameters should be chosen so that the units contributing to the conditional distribution are located near each other on the map. Then the visually observed region would also be responsible for modelling that volume in the data space. This will be addressed in further study.

References

[1] C.M. Bishop, M. Svensén, and C.K.I. Williams. GTM: The generative topographic mapping. *Neural Computation*, 10(1), 1998.

[2] Samuel Kaski. Data exploration using self-organizing maps. *A cta Polytehnic a Scandinavica, Mathematics, Computing and Management in Engineering Series No. 82*, March 1997. DTech Thesis, Helsinki University of Technology, Finland.

[3] Teuvo Kohonen. *Self-Organizing Maps*. Springer, Berlin, Heidelberg, 1995. (Second Extended Edition 1997).

[4] H. Ritter and K. Schulten. Kohonen self-organizing maps: exploring their computational capabilities. In *Proc. ICNN'88 Int. Conf. on Neural Networks*, volume I, pages 109–116, Piscataway, NJ, 1988. IEEE Service Center.

[5] A. Utsugi. Hyperparameter selection for self-organizing maps. *Neural Computation*, 9(3):623–635, 1997.