

Self-Organization with a large medical database Using the GTM for prediction & clustering

B. Rosario, D. R. Lovell, M. Niranjana, R. W. Prager and K. J. Dalton†

Cambridge University, Engineering Department,
Trumpington St., Cambridge CB2 1PZ, UK
Tel: +44 1223 332754, Fax: +44 1223 332 662,
Email: {br, drl, niranjan, rwp}@eng.cam.ac.uk

†Cambridge University Department of Obstetrics and Gynaecology,
Rosie Maternity Hospital, Robinson Way, Cambridge, CB2 2SW, UK.

Oral presentation

Abstract

We present an application of the Bishop’s GTM algorithm for the analysis of a large database of medical information and we explore a possible method to use the GTM for predictions.

The GTM algorithm finds the posterior distribution for each data point over a latent space; in this way it implements a topographic mapping from high dimensional data space into a lower dimensional latent space. Our idea is that, if the GTM preserves the intrinsic dimensionality of the data space, points “near” in the latent space should correspond to “similar” data in the data space, for instance data with similar risk. We apply the GTM algorithm to a learning set and we find for each learning point a corresponding point in the latent space. We then consider a testing set and assign to each testing point the same risk as the “closest” training point in the latent space. While we only have access to retrospective data, we already know the risk for both the learning and the testing set, so we can calculate the error obtained in this way. We show that this error is comparable with the errors found by others methods for the same data.

Keywords: Unsupervised neural network, GTM algorithm, nearest neighbourhood, risk prediction.

1 Introduction

The goal of an unsupervised neural network is to find a good representation of the probability density function from which the data were generated; this kind of statistical model allows us to make good predictions when a new data point is presented.

In this paper we use an unsupervised neural network to explore low dimensional representation of information in a 700.000 record obstetrical database.

The model we use is the *generative topographic mapping* GTM [1], an alternative to the Self-Organizing Map [2]. The interest is both in the effective analysis of the obstetric data and in the study of the model. Previous empirical

work with GTM has been restricted to a toy problem and a simulated dataset. Therefore, it seems interesting to analyse the performance of this method on a large, real-life dataset.

Moreover, we are interesting in finding new methods for predicting the risk of a bad outcome in pregnancy and this is the first unsupervised learning method used to analyse the obstetrics data.

In the next section we introduce briefly the principles of the GTM algorithm, in Section 3 we describe the data to be analysed and in Section 4 we present the results obtained with GTM to predict an adverse pregnancy outcome.

2 The GTM algorithm

The GTM algorithm [1] is an unsupervised neural network and can be considered an alternative to the Self-Organizing Map (SOM) algorithm of Kohonen [2]. Like the SOM, the GTM addresses the problem of mapping a high dimensional space into a low dimensional space; the mapping is a topographic one. Deriving from a statistical framework, the GTM overcomes some limitations of the SOM.

The GTM aims to find a representation for the distribution $p(\mathbf{t})$ of data in a D -dimensional space $\mathbf{t} = (t_1, \dots, t_D)$ in terms of a number L of “latent” variables $\mathbf{x} = (x_1, \dots, x_L)$.

Typically, the dimensionality L of the latent-variable space is less than D , since we would like to capture the fact that the data itself has an intrinsic dimensionality which is less than D and can be represented only by L variables.

We define a likelihood function $L(\mathbf{W})$, where \mathbf{W} is a weight matrix and we use a form of the expectation-maximization algorithm to maximize $L(\mathbf{W})$. Let \mathbf{W}^* be the weight matrix which maximizes the likelihood function. With the GTM algorithm we can calculate the probability distribution of \mathbf{t}^n , given \mathbf{x} and \mathbf{W}^* , $p(\mathbf{t}^n|\mathbf{x}, \mathbf{W}^*)$ and the integral $p(\mathbf{t}^n|\mathbf{W}^*)$.

Now, using Bayes’ theorem, we can calculate the posterior probability of the latent point \mathbf{x} given the data point \mathbf{t}^n and the weight matrix \mathbf{W}^*

$$p(\mathbf{x}|\mathbf{t}^n, \mathbf{W}^*) = \frac{p(\mathbf{t}^n|\mathbf{x}, \mathbf{W}^*)p(\mathbf{x})}{p(\mathbf{t}^n|\mathbf{W}^*)} \quad (1)$$

where $p(\mathbf{x})$ is the prior probability distribution over the latent space.

We can then visualize the posterior responsibility map for individual data points in latent space. It is often convenient to summarize the posterior by its mean, given for each data point \mathbf{t}^n by

$$\langle \mathbf{x}|\mathbf{t}^n, \mathbf{W}^* \rangle = \frac{\sum_{i=1}^K \mathbf{x}^i p(\mathbf{t}^n|\mathbf{x}^i, \mathbf{W}^*)}{\sum_{i=1}^K p(\mathbf{t}^n|\mathbf{x}^i, \mathbf{W}^*)}. \quad (2)$$

This calculation gives, for each point \mathbf{t}^n , a corresponding point \mathbf{x} in the latent space.

In their paper [1] the authors demonstrate the performance of the algorithm on a toy problem and on simulated data from flow diagnostics for a multi-phase oil pipeline. In the current paper, we see how GTM performs when applied to a large real-life dataset.

3 The obstetrics data

We used data from the Scottish Morbidity Register on 771,571 singleton births that occurred between 1980–91. The records were partitioned into a learning set consisting of births occurring between 1980–88, and a testing set consisting of births occurring between 1989–91.

The dataset comprises several dependent variables; in our analysis, we consider one such variable, failure to progress in labour. Put simply, failure to progress means that the natural course of labour has stalled before delivery, placing the wellbeing of both mother and child at risk.

Our ultimate aim is to determine how accurately a prediction of failure to progress could be made on the basis of the information available. We would like to develop a risk prediction system that could give an early indication of danger (*i.e.*, high risk) in order to allow prompt medical intervention.

We had to select the information relevant to the prediction of risk of this outcome. To select a relatively small set of features, we retained those features that afforded the greatest degree of discrimination between instances of failure to progress and normal labour. The criterion described in [3] resulted in the selection of 21 features shown in Table 1. We consider a 29-dimensional representation of these discrete features.

Selected predictors of failure to progress	
Predictor	Incidence
CSections	4.031%
MumAge	
NeoDxs	2.194%
Parity	
FEM STRESS INCONTINENCE	0.052%
PREM SEPARATION PLACENTA*	0.463%
ANTEPARTUM HEMORR NEC*	0.172%
THREATENED LABOR NEC*	3.087%
ANEMIA IN PREGNANCY*	6.218%
OTH CURRENT COND OF PREG*	1.841%
CEPHALIC VERSION NOS*	0.129%
TRANSVERSE/OBLIQUE LIE*	0.246%
HIGH HEAD AT TERM*	0.328%
MALPOSITION NEC*	0.482%
FETAL DISPROPORTION NOS*	0.027%
INTRAUTERINE DEATH*	0.244%
POOR FETAL GROWTH*	3.451%
EXCESSIVE FETAL GROWTH*	0.324%
OTH PLACENTAL CONDITIONS*	0.318%
GRAND MULTIPARITY*	0.079%
SUPRV HIGH-RISK PREG NEC	1.703%

Table 1: A list of the 21 predictors selected. The incidence of maternal conditions in the learning set of 540905 cases is shown in the right column. The maternal age variable **MumAge** is discretised into 7 groups: < 20, 20-24, 25-29, 30-34, 35-39, > 40, and “not known”. The variable **Parity** can take on values 0, 1, 2, 3, ≥ 4 . This discretization results in a 29 dimensional binary representation of the data.

4 Using the GTM in predicting failure to progress

The GTM operates in the principle that the high-dimensional data we measure is the result of a generative process in a much lower dimension. If the GTM can find an information preserving mapping from data space to a latent space of lower dimension, can we use the latent space representation to make good prediction about the class of new data points?

We apply the GTM algorithm to the data described in the previous section. We divide our data into a *learning set* and a *testing set*; we apply the GTM algorithm to the learning set and we find the posterior responsibility (1) and the means (2) in the latent space for each learning point. We then consider the testing set and, given the \mathbf{W}^* matrix found with GTM for the learning set, we can find *directly* the posterior responsibility (1) and the means (2) for the testing set.

The main idea is that, if GTM preserves the intrinsic dimensionality of the 29-dimensional data space, adjacent points in the latent space should correspond to “similar” data in the 29-D data space, *i.e.*, data with similar risk. We find, for each testing point, the training point which has the closest mean (2) and we assign to the testing point the risk of this training point. While we only have access to *retrospective* data, we already know the risk for both the learning and the testing set, so we can calculate the error obtained in assigning to the testing point the risk of the “nearest” (in the latent space) learning point.

We use the maximum-likelihood risk estimate

$$\frac{a_i}{a_i + b_i} \tag{3}$$

where a_i and b_i are the numbers of adverse and benign outcomes associated with the i^{th} pattern. In this situation, the cross-entropy error function ([4], [5]) is appropriate since we wish to measure the quality of our probability predictions. Assuming that y_i is the output of the network for the testing point i (*i.e.*, the risk of the nearest learning point) and that t_i is the target, *i.e.*, the effective risk of the testing point, the error function can then be written in the form

$$E = - \sum_{i=1}^N (t_i \ln y_i + (1 - t_i) \ln(1 - y_i)) \tag{4}$$

If the i^{th} pattern has a_i adverse and b_i benign outcomes, from (4) we get

$$E = - \sum_{i=1}^N (a_i \ln y_i - b_i \ln(1 - y_i)) \tag{5}$$

We then divide the error (5) by the total number of cases in the testing set in order to obtain the Average Negative Log-Likelihood. This number can be compared to the same Avg.Neg.LL found by other methods for the same data. The results are shown in Table 2.

We can see the GTM performs a good mapping from a high dimensional input space to a low dimensional latent space; we lose few information considering a 2 (or even 1) dimensional latent space. This experiment is one of the first experiments to propose a sensible way to use the GTM for predictions.

System	Avg.Neg.LL
Limit attainable on test set	0.2691
Ensemble of networks	0.2748
Nearest neighbour in the data space	0.2753
GTM 2D (Nearest neighbour in 2-D latent space)	0.2756
Hierarchical Dirichlet model	0.2756
GTM 1D (Nearest neighbour in 1-D latent space)	0.2757
Ensemble of direct weight networks	0.2757
Ensemble of logistic regressors	0.2759
Logistic regression	0.2759

Table 2: Comparison of the results obtained with various methods. The method to obtain the limit attainable on test set is described in [3].

5 Using the GTM to detect clusters in the data

The Self-Organizing Map (SOM) algorithm of Kohonen [2] which inspired the GTM is often regarded as defining a projection from the D -dimensional data space onto a two-dimensional ‘feature’ space, which allows applications to problems in data visualization. For the purposes of data visualization, the GTM algorithm finds a full posterior distribution in the latent space. The unsupervised methods are often especially used to visualize data, in order to detect possible clusters in the data set (see the experimental results in [1]).

We also tried to visualize in the latent space possible clusters of the data. Figure shows 2 some examples of the the risk plotted in function of the latent space. We would like to see if GTM could locate clusters of high risk patient but the clusters we obtained were not very clear. It seems difficult to visualize any clusters; the reason might be the high dimension of the data space or the difficulty to interpret clustering but for the purpose of risk prediction our interest is classification not visualization.

6 Summary

We wanted to explore the idea that a low dimensional projection of the data could be useful in making prediction of bad outcome in pregnancy. We used the GTM algorithm to make such projection and found that the mapping performed by GTM preserves the information very well. Having data in a 29 dimensional space, and projecting them into a 2 or 1 dimensional latent space, we lose very few information.

References

- [1] C. M. Bishop, M. Svensen, and C. K. I. Williams, “Gtm: A principled alternative to the self-organizing map,” *Submitted to Neural Computation*, 1996.
- [2] T. Kohonen, “Self-organized formation of topologically correct feature maps,” *Biological Cybernetics*, vol. 43, pp. 59–69, 1982.

- [3] D. R. Lovell, B. Rosario, M. Niranjana, R. W. Prager, K. J. Dalton, and R. Derom, "Design, construction and evaluation of systems to predict risk in obstetrics," vol. Submitted to International Journal of Biomedical Computing, 1997.
- [4] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.
- [5] J. J. Hopfield, "Learning algorithms and probability distributions in feed-forward and feed-back networks," *Proceedings of the National Academy of Sciences*, vol. 84, pp. 8429–8433, 1987.

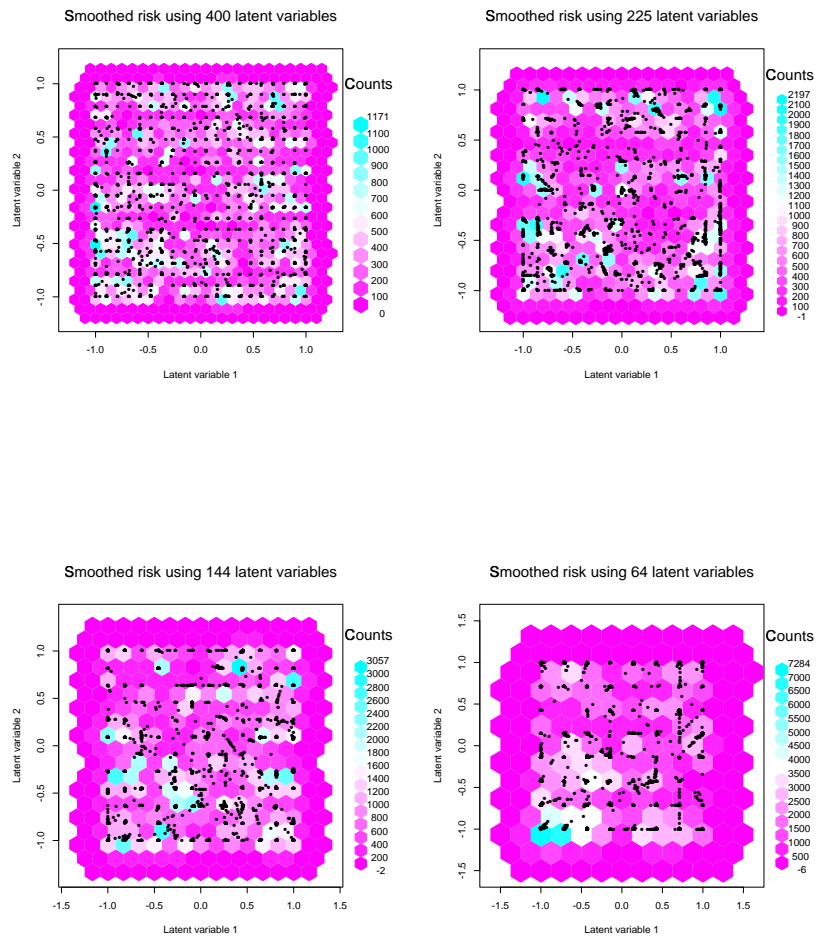


Figure 1: The risk in the latent space for different values of the number of the latent points for a 2-D latent space. The points are the means of the data in the latent space; the colour is function of the risk.