ESANN'2005 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 27-29 April 2005, d-side publi., ISBN 2-930307-05-6.

Handling outliers and missing data in brain tumor clinical assessment usign *t*-GTM

Alfredo Vellido[1], Paulo J.G. Lisboa[2], and Dolores Vicente[1]

[1]Universitat Politècnica de Catalunya (UPC). Soft Computing Group.
C. Jordi Girona, 1-3, 08034, Barcelona, Spain
[2]Liverpool John Moores University (LJMU). Neural Computation Group.
Byrom St, L3 3AF, Liverpool, U.K

Abstract. Uncertainty is inherent to medical decision making, and automated decision support systems should aim to reduce it. In this paper, MR spectral data are considered in a problem of discrimination of brain tumour types and grades. Models to fit these data can be affected by two sources of uncertainty that might occur in the data: the presence of outliers and data incompleteness. A model for multivariate data clustering and visualization, the GTM, is here redefined as a mixture of Student *t*-distributions that is robust towards outliers while providing missing values imputation. The effectiveness of this model on the MRS data is demonstrated empirically.

1 Introduction

The reduction of uncertainty must be the primary goal of any automated system designed to provide support for medical diagnostic and prognostic decision making. In this brief paper, we deal with the decision problem of brain tumour discrimination through Magnetic Resonance Spectroscopy (MRS) information, obtained from living tissue. In practice, it is not unusual that expert assessments of the type and grade of the tumour be made on the basis of visual inspection of MR spectra and prior experience. Robust decision support systems are of paramount importance in these circumstances.

Two potential sources of uncertainty in diagnosis or prognosis based on MR spectrometric data are the presence of outliers and the incompleteness of the available data in the form of missing values. The decision support model discussed in this paper is a variation on the standard Generative Topographic Mapping (GTM:[1]). The GTM allows for the simultaneous clustering and visualization of multivariate data. It was originally described as an alternative to the neural network-inspired Self-Organizing Maps (SOM:[2]) with sound probabilistic foundations. The GTM can also be seen as a constrained mixture of distributions. This definition as a constrained model makes it less flexible than general mixtures, but the renounce to full flexibility is compensated by its data visualization capabilities.

The GTM was originally defined as a constrained mixture of Gaussians. It is well reported [3] that Gaussian mixture models lack robustness in the presence of outlier observations in the data sample. Several recent studies [4,5,6] have suggested the use of multivariate Student *t*-distributions as a robust alternative to Gaussians for mixture models. Here, we redefine the GTM as a constrained mixture of Student *t*-distributions, the *t*-GTM, and show its ability to successfully identify data outliers

and, simultaneously, minimize their negative impact on the calculation of the model parameters. The management of the uncertainty introduced by data incompleteness is our second goal. In the following sections, details are provided on how to integrate missing data imputation as part of the *t*-GTM model fitting to the data. The resulting model plays a double role: it deals robustly with outliers while simultaneously imputes missing values, allowing the exploration of multivariate data through visualization at a reasonable computational cost.

2 Generative Topographic Mapping as a mixture of *t*-distributions

The GTM is a non-linear latent variable model that defines a mapping from a low dimensional latent space onto the multivariate data space. The mapping is carried through by a set of basis functions generating a (mixture) density distribution, and it is defined as a generalized linear regression model:

$$\mathbf{y} = \boldsymbol{\Phi}(\mathbf{u})\mathbf{W} \tag{1}$$

where $\boldsymbol{\Phi}$ is a set of M basis functions $\boldsymbol{\Phi}(\mathbf{u}) = (\phi_1(\mathbf{u}),...,\phi_M(\mathbf{u}))$. These basis functions were originally defined as spherically symmetric Gaussians $\phi_m(\mathbf{u}) = exp\left\{-\left\|\mathbf{u}-\mu_m\right\|^2 \Big/ 2\sigma^2\right\}$ to deal with continuous data, with μ_m the centres of the basis functions and σ their common width; $\mathbf{W}$ is a matrix of adaptive weights w_{md} that defines the mapping, and $\mathbf{u}$ is a point in latent space. This latent space can be discretized as a regular grid of K latent points $\mathbf{u}_k$, similar to that of the SOM. A probability distribution for the data can then be defined, leading to an expression for the complete log-likelihood $L_c(\mathbf{W},\beta|\mathbf{X})$. The Expectation-Maximization (E-M) algorithm can be used to obtain the Maximum Likelihood (ML) estimates of the adaptive parameters $\mathbf{W}$ and β. Details of this procedure can be found in [1].

For the Gaussian GTM, the presence of outliers is likely to negatively bias the estimation of parameters $\mathbf{W}$ and β, and it is also likely to result in extreme estimates of the posterior probabilities of component membership [3]. To overcome this limitation, the GTM is here redefined as a constrained mixture of Student *t*-distributions: the *t*-GTM. Assuming now that the basis functions $\boldsymbol{\Phi}$ are Student *t*-distributions, the data probability can be defined as

$$P(\mathbf{x}|\mathbf{u},\mathbf{W},\beta,\nu) = \frac{\Gamma\left(\nu/2 + D/2\right)\beta^{D/2}}{\Gamma\left(\nu/2\right)(\nu\pi)^{D/2}}\left(1+\beta/\nu\left\|\mathbf{y}-\mathbf{x}\right\|^2\right)^{-\frac{\nu+D}{2}} \tag{2}$$

where $\boldsymbol{\Gamma}(\cdot)$ is the gamma function and the parameter ν can be understood as a tuner that adapts the level of robustness (divergence from normality) for the mixture. This leads to a new complete log-likelihood:

$$L_c(\mathbf{W},\beta,\nu|\mathbf{X})=\sum_{n=1}^{N} log\left\{\frac{1}{K}\sum_{k=1}^{K}\frac{\Gamma\left(\nu_k/2+D/2\right)\beta^{D/2}}{\Gamma\left(\nu_k/2\right)(\nu_k\pi)^{D/2}}\left(1+\beta/\nu_k\left\|\mathbf{y}_k-\mathbf{x}_n\right\|^2\right)^{-\frac{\nu_k+D}{2}}\right\}. \quad (3)$$

From this expression, ML estimates of the adaptive parameters $\mathbf{W}$ and β can be calculated, using the E-M algorithm. Further details can be obtained from [7]. A similar closed expression for parameter ν cannot readily be obtained, although an adequate value can be calculated through preliminary runs of the algorithm.

Missing data imputation through t-GTM

Missing data imputation arises naturally as part of the ML estimation of the *t*-GTM parameters via de E-M algorithm [8]. Following [9], two separate submatrices: $\mathbf{X}^o$, consisting of the observed data, and $\mathbf{X}^m$, consisting of the missing data, are considered. The E-step of the E-M algorithm includes the calculation of the expected complete log-likelihood. The definition of submatrices $\mathbf{X}^o$ and $\mathbf{X}^m$ entails a modification of (3) that becomes:

$$L_c(\mathbf{W},\beta,\nu|\mathbf{X}^o,\mathbf{X}^m,\mathbf{Z})=\sum_{n=1}^{N}\sum_{k=1}^{K} z_{kn}\, log\left\{C_k\left[1+\beta/\nu_k\left(\left\|\mathbf{y}_k^o-\mathbf{x}_n^o\right\|^2+\left\|\mathbf{y}_k^m-\mathbf{x}_n^m\right\|^2\right)\right]^{-\frac{\nu_k+D}{2}}\right\}, \quad (4)$$

where C_k is a summary coefficient and $\mathbf{Z}$ is an indicator matrix, with elements z_{kn} describing our lack of knowledge of which latent point $\mathbf{u}_k$ is responsible for the generation of data point $\mathbf{x}_n$. The sufficient statistics that must be calculated prior to the M-step of the E-M algorithm are: the expected values of the unknown z_{kn}, $E\left[z_{kn}|\mathbf{x}_n^o,\mathbf{W},\beta,\nu_k\right]=P(k|\mathbf{x}_n,\mathbf{W},\beta,\nu)$, calculated using only the observed data, and the interactions between z_{kn} and the first and second moments of $\mathbf{x}_n^m$, $E\left[z_{kn}\mathbf{x}_n^m|\mathbf{x}_n^o,\mathbf{W},\beta,\nu\right]$ and $E\left[z_{kn}\mathbf{x}_n^m\mathbf{x}_n^{mT}|\mathbf{x}_n^o,\mathbf{W},\beta,\nu\right]$. We first define [8,9] the expectation:

$$E\left[\mathbf{x}_n^m|z_{kn}=1,\mathbf{x}_n^o,\mathbf{W},\beta,\nu_k\right]=\hat{\mathbf{x}}_{kn}^m=\left(\mathbf{y}_k^m\right)^{old}, \quad (5)$$

which leads to the calculation of both $E\left[z_{kn}\mathbf{x}_n^m|\mathbf{x}_n^o,\mathbf{W},\beta,\nu_k\right]=\hat{z}_{kn}\hat{\mathbf{x}}_{kn}^m$ and $E\left[z_{kn}\mathbf{x}_n^m\mathbf{x}_n^{mT}|\mathbf{x}_n^o,\mathbf{W},\beta,\nu_k\right]=\hat{z}_{kn}\left(\left(\beta^{-1}\right)^{old}+\hat{\mathbf{x}}_{kn}^{mT}\hat{\mathbf{x}}_{kn}^m\right)$, where $\hat{z}_{kn}$ is given by:

$$\hat{z}_{kn}=P(k|\mathbf{x}_n,\mathbf{W},\beta,\nu_k)=\frac{C_k\left(1+\beta/\nu_k\left\|\mathbf{y}_k^o-\mathbf{x}_n^o\right\|^2\right)^{-\frac{\nu_k+D}{2}}}{\sum_{k'=1}^{K}C_{k'}\left(1+\beta/\nu_{k'}\left\|\mathbf{y}_{k'}^o-\mathbf{x}_n^o\right\|^2\right)^{-\frac{\nu_{k'}+D}{2}}} \quad (6)$$

and *old* stands for previous iterations. The missing data imputation is now straightforward, performed according to:

$$E\left[\mathbf{x}_n^m \middle| \mathbf{x}_n^o, \mathbf{W}, \beta, \nu_k\right] = \sum_{k=1}^{K} \hat{z}_{kn} E\left[\mathbf{x}_n^m \middle| z_{kn} = 1, \mathbf{x}_n^o, \mathbf{W}, \beta, \nu_k\right] = \sum_{k=1}^{K} \hat{z}_{kn} \left(\mathbf{y}_k^m\right)^{old} \quad (7)$$

This imputation procedure completes the data and would allow their full visualization and clustering on the low-dimensional latent space (although in this paper the focus is placed on the assessment of the model's ability to handle outliers and missing data imputation). Update expressions for parameters $\mathbf{W}$ and β can be calculated in closed form in the M-step of the E-M algorithm, using those now reconstructed data. Details are omitted for the sake of brevity.

3 MRS and brain tumour data

MRS is a non-invasive tool capable of providing a detailed fingerprint of the biochemistry of living tissue. Diagnosis and prognosis based on Magnetic Resonance Imaging (MRI) can sometimes be uncertain. The additional information contained in the MR spectrum can help the clinical expert by disambiguating decisions. The data used in this study consist of 98 single voxel PROBE (PROton Brain Exam system) spectra acquired *in vivo* for five viable tumour types (Astrocytes, Glioblastomas, Metastases, Meningiomas, and Oligodendrogliomas) and cystic regions from tumours that, given their specific composition, are likely to differ from the tumours themselves. A description of the automated protocol used for data acquisition can be found in [10]. The spectra were digitised, sampling the region known to contain clinically relevant metabolic information, into 194 frequency intensity values. The high dimensionality of the problem makes either feature extraction or variable selection necessary. In [10], a process based on Multivariate Bayesian Variable Selection was shown to provide a good description of the data set in the form of 6 frequency intensities, corresponding to Fatty Acids, Lactate, a compound-unassigned peak, Glutamine, Choline, and Taurine-Inositol. These 6 variables will be the inputs to the *t*-GTM model.

4 Experimental results

According to [3], a given data instance could be considered as outlier if the value of:

$$O_n = \sum_k \hat{z}_{kn} \frac{\nu + D}{\nu + \beta \left\| \mathbf{y}_k - \mathbf{x}_n \right\|^2} \quad (8)$$

was sufficiently small or, equivalently, sufficiently large the value of

$$O_n^* = \sum_k \hat{z}_{kn} \beta \left\| \mathbf{y}_k - \mathbf{x}_n \right\|^2 \quad (9)$$

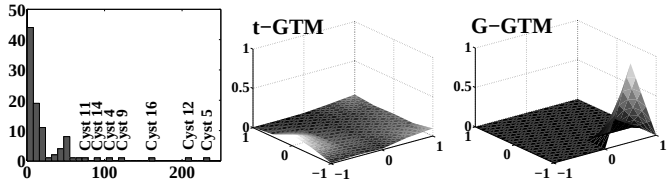

Figure 1: (Left): Histogram of the statistic (9); outliers are characterised by large values that, in this case, mostly correspond to cystic region spectra. (Centre): Posterior probability of all *t*-GTM mixture components in the square latent grid, given outlier spectrum cystic region 5. (Right): Posterior probability of all Gaussian GTM components, given the same outlier.

The histogram in Figure 1 (Left) reflects the values of the statistic (9) for the brain tumour data, providing the decision maker with a measure of *novelty* that might be used to focus attention on MR spectra that do not fit the main distributions estimated by the model. The 7 data instances with largest values for statistic (9) are cystic regions. In more detail, 14 out the 17 cystic regions in the data set are included in the three highest decile intervals of (9). This is consistent with the specificity of these regions, as mentioned in the previous section. Figure 1 (Centre, Right) displays the posterior probability of all latent points (mixture components) in the GTM square grid, given a data point, for an example spectrum which was labelled as a extreme outlier according to (9). Clearly, no mixture component of the *t*-GTM takes main *responsibility* for this spectrum, illustrating how the *t*-GTM effectively minimizes the negative impact of outliers on the modelling of the distributions. In contrast, one single component of the Gaussian GTM takes most responsibility for the outlier.

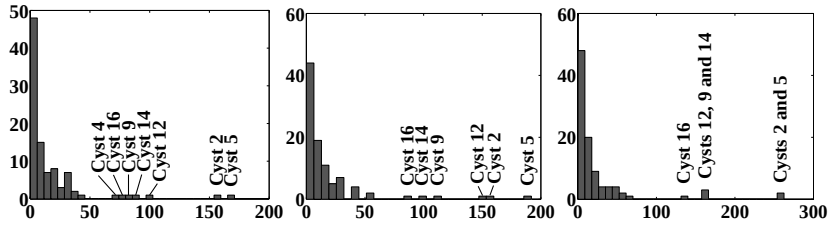

Figure 2: Histograms of the statistic (9) at different levels of data incompleteness (5, 10, and 20%, in turn, from left to right). In order to allow comparison with Fig.1 (Left), the most extreme cystic region outliers have been labelled.

The attention is now turned towards the effect of data incompleteness on outlier identification by the model. The proposed method of missing data imputation using the *t*-GTM is applied to three levels of incompleteness. The corresponding histograms of the statistic (9), in Figure 2, show that, even at rather high levels of incompleteness, the imputation procedure still completes the data in such a way that almost the same cystic regions are singled out as the most extreme outliers.

5 Conclusion

Artificial Intelligence models in general and Artificial Neural Networks in particular have a longstanding track record as tools for automated decision support in medical applications concerning diagnosis and prognosis [11]. MRI is one field in which prognostic decision making entails expert subjectivity [12]. The analysis of MRS data has a considerable potential as a tool to support decision making, but data can be also a source of uncertainty in the form of outlier presence or data incompleteness.

In this paper we have introduced a redefinition of GTM as a constrained mixture of *t*-distributions. It has been shown to behave robustly in the presence of outliers, while minimizing, through imputation, the negative effect of data incompleteness.

Acknowledgements

The authors gratefully acknowledge C. Arús from the Universitat Autónoma de Barcelona for making available the data for this study.

References

[1] C.M. Bishop, M. Svensén and C.K.I. Williams, GTM: The Generative Topographic Mapping, *Neural Computation*, 10:215-234, 1998.

[2] T. Kohonen. *Self-organizing Maps (3rd ed.)*, Springer-Verlag, Berlin, 2000.

[3] D. Peel and G.J. McLachlan, Robust mixture modelling using the t distribution, *Statistics and Computing*, 10:339–348, 2000.

[4] C. Archambeau, F. Vrins and M. Verleysen, Flexible and robust Bayesian classification by finite mixture models. In M. Verleysen, editor, *proceedings of the 12th European Symposium on Artificial Neural Networks* (ESANN 2004), D-Side Pub., pages 75-80, Bruges (Belgium), 2004.

[5] C.M. Bishop and M. Svensén: Robust Bayesian mixture modelling. In M. Verleysen, editor, *proceedings of the 12th European Symposium on Artificial Neural Networks* (ESANN 2004), D-Side Pub., pages 69-74, Bruges (Belgium), 2004.

[6] H.X. Wang, Q.B. Zhang, B. Luo and S. Wei, Robust mixture modelling using multivariate t-distribution with missing information, *Pattern Recognition Letters*, 25:701–710, 2004.

[7] A. Vellido. Generative Topographic Mapping as a constrained mixture of Student *t*-distributions: Theoretical developments, Technical Report LSI-44-7-R, Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, 2004.

[8] Z. Ghahramani and M.I. Jordan. Learning from incomplete data. Technical Report, AI Laboratory, MIT, MA, US, 1994.

[9] Y. Sun, P. Tiňo and I. Nabney. GTM-based data visualization with incomplete data. Technical Report, NCRG, Aston University, Birmingham, England, 2001.

[10] Y. Huang, P.J.G. Lisboa and W. El-Deredy, Tumour grading from Magnetic Resonance Spectroscopy: A comparison of feature extraction with variable selection, *Statistics in Medicine,* 22:147-164, 2003.

[11] P.J.G. Lisboa, A review of evidence of health benefit from artificial neural networks in medical intervention, *Neural Networks*, 15:11-39, 2002.

[12] P.J.G. Lisboa, A. Vellido and H. Wong, Outstanding issues for clinical decision support with Neural Networks. In H. Malmgren, M. Borga and L. Niklasson , editors, *Artificial Neural Networks in Medicine and Biology*, pages 63-71, Springer, London, 2000.