TWO-STAGE CLUSTERING OF A HUMAN BRAIN TUMOUR DATASET USING MANIFOLD LEARNING MODELS[*]

Raúl Cruz-Barbosa and Alfredo Vellido
Dept. de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya
Edifici Omega, Campus Nord, C. Jordi Girona, 1-3, Barcelona, 08034, Spain

Keywords: Brain tumours, MRS, Generative Topographic Mapping, two-stage clustering, outliers.

Abstract: This paper analyzes, through clustering and visualization, Magnetic Resonance spectra of a complex multi-center human brain tumour dataset. Clustering is performed as a two-stage process, in which the models used in the first stage are variants of Generative Topographic Mapping (GTM). Class information-enriched variants of GTM are used to obtain a primary cluster description of the data. The number of clusters used by GTM is usually large and does not necessarily correspond to the overall class structure. Consequently, in a second stage, clusters are agglomerated using K-means with different initialization strategies, some of them defined *ad hoc* for the GTM models. We evaluate if the use of class information influence the brain tumour cluster-wise class separability resulting from the process. We also resort to a robust variant of GTM that detects outliers while effectively minimizing their negative impact in the clustering process.

1 INTRODUCTION

Medical decision making is usually riddled with uncertainty, especially in sensitive settings such as non-invasive brain tumour diagnosis. The brain tumour data analysed in this study are obtained by Magnetic Resonance Spectroscopy (MRS). Information derived from the MR spectra can contribute to the evidence base available for a particular patient, providing support to clinicians.

The fields of Machine Learning and Statistics coexist with data analysis as a common target. An example can be found in Finite Mixture Models (Figueiredo and Jain, 2002). In practical scenarios, such as medical decision making, these models could benefit from data visualization. Finite Mixture Models can be endowed with visualization capabilities provided certain constrains are enforced, such as forcing the mixture components to be centred in a low-dimensional manifold embedded in the observed data space, as in Generative Topographic Mapping (GTM) (Bishop et al., 1998), which can be seen as a probabilistic alternative to Self-Organizing Maps (SOM) (Kohonen, 1995) for data clustering and visualization. When available class information can also be integrated as part of the GTM training to enrich the cluster structure definition (Cruz and Vellido, 2006). The resulting models will be used in our experiments to analyze a complex MRS dataset.

GTM-based models do not place any strong restriction on the number of mixture components (or clusters), in order to achieve an appropriate visualization of the data. This richly detailed cluster structure does not necessarily match the more global cluster and class structures of the data. In this scenario, a two-stage clustering procedure may be useful to uncover such global structure (Vesanto and Alhoniemi, 2000). GTM and its variants can be used in the first stage to generate a detailed cluster partition in the form of a mixture of components. The centres of these components can be further clustered in the second stage. For that role, the well-known K-means algorithm is used in this study.

The first goal of the paper is assessing to what extent the introduction of class information improves the final cluster-wise class separation. The issue remains of how we should initialize K-means in the sec-

[*] Alfredo Vellido is a Ramón y Cajal researcher and acknowledges funding from the MEC project TIN2006-08114. Raúl Cruz-Barbosa acknowledges SEP-SESIC (PROMEP program) of México for his PhD grant. Authors gratefully acknowledge the former INTERPRET project centres (http://azizu.uab.es/INTERPRET) for making available and managing the data for this study.

Cruz-Barbosa R. and Vellido A. (2008).
TWO-STAGE CLUSTERING OF A HUMAN BRAIN TUMOUR DATASET USING MANIFOLD LEARNING MODELS.
In *Proceedings of the First International Conference on Bio-inspired Systems and Signal Processing*, pages 191-196
DOI: 10.5220/0001058501910196

ond clustering stage. Random initialization (Vesanto and Alhoniemi, 2000) does not make use of the prior knowledge generated in the first stage of the procedure and requires a somehow exhaustive search of the initialization space. Here, we propose two different ways of introducing such prior knowledge as fixed initialization. These procedures, resulting from GTM properties, allow significant computational savings.

In section 2, we summarily introduce the GTM and its *t*-GTM and class-enriched variants, as well as the two-stage clustering procedure with its alternative initialization strategies. Several experimental results are provided and discussed in section 3, while a final section outlines some conclusions.

2 TWO-STAGE CLUSTERING

2.1 The GTM Models

The standard GTM is a non-linear latent variable model defined as 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 constrained mixture density distribution. It is defined as a generalized linear regression model:

$$\mathbf{y} = \phi(\mathbf{u})\mathbf{W}, \tag{1}$$

where ϕ are M basis functions $\phi(\mathbf{u}) = (\phi_1(\mathbf{u}), ..., \phi_M(\mathbf{u}))$. For continuous data of dimension D, spherically symmetric Gaussians are an obvious choice of basis function; $\mathbf{W}$ is a matrix of adaptive weights w_{md} that defines the mapping, and $\mathbf{u}$ is a point in latent space. To avoid computational intractability a regular grid of K points $\mathbf{u}_k$ can be sampled from the latent space. Each of them, which can be considered as the representative of a data cluster, has a fixed prior probability $p(\mathbf{u}_k) = 1/K$ and is mapped, using (1), into a low dimensional manifold non-linearly embedded in the data space. This latent space grid is similar in design and purpose to that of the visualization space of the SOM. A probability distribution for the multivariate data $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^{N}$ can then be defined, leading to the following expression for the log-likelihood:

$$L = \sum_{n=1}^{N} \ln \left\{ \frac{1}{K} \sum_{k=1}^{K} \left(\frac{\beta}{2\pi}\right)^{D/2} \exp\left\{ \frac{-\beta\|\mathbf{y}_k - \mathbf{x}_n\|^2}{2} \right\} \right\} \tag{2}$$

where $\mathbf{y}_k$, usually known as *reference* or *prototype* vectors, are obtained for each $\mathbf{u}_k$ using (1); and β is the inverse of the noise model variance. The EM algorithm is an straightforward alternative to obtain the Maximum Likelihood (ML) estimates of the adaptive parameters of the model, namely $\mathbf{W}$ and β.

The class-GTM model is an extension of GTM that makes use of the available class information. The main goal of this extension is to improve class separability in the clustering results of GTM. For the Gaussian version of the GTM model (Sun et al., 2002; Cruz and Vellido, 2006), this entails the calculation of the posterior probability of a cluster representative $\mathbf{u}_k$ given the data point $\mathbf{x}_n$ and its class label c_n, or class-conditional *responsibility* $\hat{z}_{kn}^{c} = p(\mathbf{u}_k|\mathbf{x}_n, c_n)$, as part of the E step of the EM algorithm. It can be calculated as:

$$\begin{aligned} \hat{z}_{kn}^{c} &= \frac{p(\mathbf{x}_n,c_n|\mathbf{u}_k)}{\sum_{k'=1}^{K} p(\mathbf{x}_n,c_n|\mathbf{u}_{k'})} \\ &= \frac{p(\mathbf{x}_n|\mathbf{u}_k)p(c_n|\mathbf{u}_k)}{\sum_{k'=1}^{K} p(\mathbf{x}_n|\mathbf{u}_{k'})p(c_n|\mathbf{u}_{k'})} \\ &= \frac{p(\mathbf{x}_n|\mathbf{u}_k)p(\mathbf{u}_k|c_n)}{\sum_{k'=1}^{K} p(\mathbf{x}_n|\mathbf{u}_{k'})p(\mathbf{u}_{k'}|c_n)}, \end{aligned} \tag{3}$$

and, being T_i each class,

$$p(\mathbf{u}_k|T_i) = \frac{\frac{\sum_{n;c_n=T_i} p(\mathbf{x}_n|\mathbf{u}_k)}{\sum_n p(\mathbf{x}_n|\mathbf{u}_k)}}{\frac{\sum_{k'} \sum_{n;c_n=T_i} p(\mathbf{x}_n|\mathbf{u}_{k'})}{\sum_n p(\mathbf{x}_n|\mathbf{u}_{k'})}} \tag{4}$$

The rest of the model's parameters are estimated following the standard EM procedure.

For the Gaussian GTM, the presence of outliers is likely to negatively bias the estimation of the adaptive parameters, distorting the clustering results. In order to overcome this limitation, the GTM was recently redefined (Vellido, 2006; Vellido and Lisboa, 2006) as a constrained mixture of Student's *t* distributions: the *t*-GTM, aiming to increase the robustness of the model towards outliers. The mapping described by Equation (1) remains, with the basis functions now being Student's *t* distributions and leading to the definition of the following mixture density:

$$\begin{aligned} &p(\mathbf{x}|\mathbf{W},\beta,\nu_k) = \\ &\frac{1}{K} \sum_{k=1}^{K} \frac{\Gamma(\frac{\nu_k+D}{2})\beta^{D/2}}{\Gamma(\frac{\nu_k}{2})(\nu_k\pi)^{D/2}} \left(1 + \frac{\beta}{\nu_k}\|\mathbf{y}_k - \mathbf{x}_n\|^2\right)^{\frac{\nu_k+D}{2}} \end{aligned} \tag{5}$$

where $\Gamma(\cdot)$ is the gamma function and the parameter $\nu = (\nu_1, \ldots, \nu_K)$ represents the degrees of freedom for each component k of the mixture, so that it can be viewed as a tuner that adapts the level of robustness (divergence from normality) for each component. This density leads to the redefinition of the model log-likelihood and, again, the estimation of the corresponding adaptive parameters using EM. The extension to class-*t*-GTM is straightforward and is omitted here for the sake of brevity.

2.2 Two-Stage Clustering based on GTM

In the first stage of the proposed two-stage clustering procedure, the GTM models are trained to obtain the representative prototypes (detailed clustering) of the observed dataset. In this study, the resulting prototypes $\mathbf{y}_k$ of the GTM models are further clustered using the K-means algorithm. In a similar two-stage procedure to the one described in (Vesanto and Alhoniemi, 2000), based on SOM, the second stage K-means initialization in this study is first randomly replicated 100 times, subsequently choosing the best available result, which is the one that minimizes the error function $E = \sum_{c=1}^{C} \sum_{\mathbf{x} \in G_c} \|\mathbf{x} - \mu_c\|^2$, where C is the final number of clusters in the second stage and μ_c is the centre of the K-means cluster G_c. This approach seems somehow wasteful, though, as the use of GTM instead of SOM can provide us with richer a priori information to be used for fixing the K-means initialization in the second stage.

Two novel fixed initialization strategies that use the prior knowledge obtained by GTM in the first stage are proposed. They are based on the Magnification Factors (MF) and the Cumulative Responsibility (CR). The MF measure the level of stretching that the mapping undergoes from the latent to the data spaces. Areas of low data density correspond to high distorsions of the mapping (high MF), whereas areas of high data density correspond to low MF. The MF is described in terms of the derivatives of the basis functions $\phi_j(\mathbf{u})$ in the form:

$$\frac{dA'}{dA} = \det{}^{1/2}\left(\psi^{\mathbf{T}}\mathbf{W}^{\mathbf{T}}\mathbf{W}\psi\right), \qquad (6)$$

where ψ has elements $\psi_{ji} = \partial\phi_j / \partial u^i$ (Bishop et al., 1997) and dA' and dA are, in turn, infinitesimal rectangles in the manifold and latent spaces. If we choose C to be the final number of clusters for K-means in the second stage, the first proposed fixed initialization strategy will consist on the selection of the class-GTM prototypes corresponding to the C non-contiguous latent points with lowest MF for K-means initialization. That way, the second stage algorithm is meant to start from the areas of highest data density.

The CR is the sum of responsibilities over all data points in $\mathbf{X}$ for each cluster k:

$$CR_k = \sum_{n=1}^{N} \hat{z}^{c}_{kn} \qquad (7)$$

The second proposed fixed initialization strategy, based on CR, is similar in spirit to that based on MF. Again, if we choose C to be the final number of clusters for K-means in the second stage, the fixed initialization strategy will now consist on the selection of the GTM prototypes corresponding to the C non-contiguous latent points with highest CR. That is, the second stage is meant to start from those prototypes that are found in the first stage to be most responsible for the generation of the observed data.

3 EXPERIMENTS

3.1 Human Brain Tumour Data

The multi-center data used in this study consists of 217 single voxel PROBE (PROton Brain Exam system) MR spectra acquired in vivo for six brain tumour types: meningiomas (58 cases), glioblastomas (86), metastases (38), astrocytomas (22), oligoastrocytomas (6), and oligodendrogliomas (7). For the analyses, the spectra were grouped into three types (typology that will be used in this study as class information), as in (Tate et al., 2006): high grade malignant (metastases and glioblastomas), low grade gliomas (astrocytomas, oligodendrogliomas and oligoastrocytomas) and meningiomas. The clinically relevant regions of the spectra were sampled to obtain 200 frequency intensity values. The high dimensionality of the problem was compounded by the small number of spectra available, which is commonplace in MRS data analysis.

3.2 Experimental Design and Settings

The GTM, *t*-GTM and their class-enriched counterparts were implemented in MATLAB®. For the experiments reported next, the adaptive matrix $\mathbf{W}$ was initialized, following a PCA-based procedure described in (Bishop et al., 1998). This ensures the replicability of the results. The grid of latent points $\mathbf{u}_k$ was fixed to a square 20x20 layout for the MRS dataset. The corresponding grid of basis functions ϕ was equally fixed to a 5x5 square layout.

The goals of these experiments are fourfold. First, we aim to assess whether the inclusion of class information in the first stage of the procedure results in any improvement in terms of cluster-wise class separability (and under what circumstances) compared to the procedure using standard GTM. Second, we aim to assess whether the two-stage procedure improves, in the same terms, on the use of direct clustering of the data using K-means. Third, we aim to test whether the second stage initialization procedures based on MF and the CR of the class-GTM, described in section 2.2, retain the cluster-wise class separability capabilities of the two-stage clustering procedure in which K-means is randomly initialized. In fourth place, we aim to explore the properties of the structure of the dataset concerning atypical data. For that, we use the *t*-GTM (Vellido, 2006), as described in section 2.1.

The clustering results of all models will be first compared visually, which should help to illustrate the visualization capabilities of the models. Beyond the visual exploration, the second stage clustering results

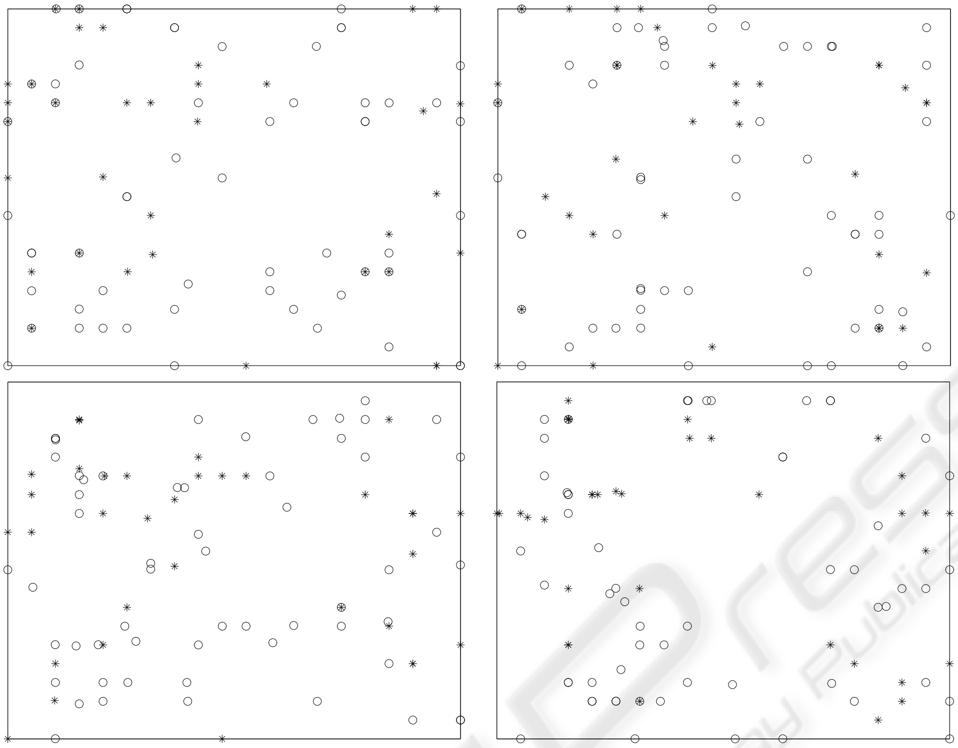

Figure 1: Representation, on the 2-dimensional latent space of GTM and its variants, of a part of the tumour dataset. It is based on the mean posterior distributions for the data points belonging to low grade gliomas ('*') and meningiomas ('o'). The axes of the plot convey no meaning by themselves and are kept unlabeled. (Top left): GTM without class information. (Top right): class-GTM. (Bottom left): *t*-GTM without class information. (Bottom right): class-*t*-GTM.

should be explicitly quantified in terms of cluster-wise class separability. For that purpose, the following entropy-like measure is proposed:

$$E_{G_c}(\{T_i\}) = -\sum_{\{G_c\}} P(G_c) \sum_{\{T_i\}} P(T_i|G_c)\ln P(T_i|G_c) \qquad (8)$$
$$= -\sum_{c=1}^{C} \frac{K_{G_c}}{K} \sum_{i=1}^{|\{T_i\}|} p_{ci} \ln p_{ci}$$

Sums are performed over the set of classes (tumour types) $\{T_i\}$ and the K-means clusters $\{G_c\}$; K is the total number of prototypes; K_{G_c} is the number of prototypes assigned to the c^{th} cluster; $p_{ci} = \frac{K_{G_{c}i}}{K_{G_c}}$, where $K_{G_{c}i}$ is the number of prototypes from class i assigned to cluster c; and, finally, $|\{T_i\}|$ is the cardinality of the set of classes. An entropy of 0 corresponds to the case of no clusters being assigned prototypes corresponding to more than one class.

Given that the use of a second stage in the clustering procedure is intended to provide final clusters that best reflect the overall structure of the data, the problem remains of what is the most adequate number of clusters. In this paper we do not use any cluster validity index and we just evaluate the entropy measure for solutions from 2 up to 10 clusters.

3.3 Results and Discussion

In the first stage of the two-stage clustering procedure, GTM, *t*-GTM and their class-enriched variants class-GTM and class-*t*-GTM were trained to model the human brain tumour dataset. The resulting prototypes $\mathbf{y}_k$ were then clustered in the second stage using the K-means algorithm. This last stage was performed with three different initializations, as described in section 2.2. In all cases, K-means was forced to yield a given number of final clusters, from 2 up to 10. The entropy was calculated for all settings.

Before considering the entropy results, visualization maps (obtained using the mean of the posterior distribution: $\sum_{k=1}^{K} \mathbf{u}_k \hat{z}_{kn}$ or $\sum_{k=1}^{K} \mathbf{u}_k \hat{z}_{kn}^{c}$) of all the trained models in the first stage were generated. Three hypotheses were made for the clustering results visualized here. First, the use of class information in the clustering models should yield visualization maps where classes are separated better than in those models which do not use it. Second, the use of *t*-

GTM should help to diminish the influence of outliers and the visualization maps generated with these models should show the data more homogeneously distributed throughout the visualization maps than in Gaussian GTM, which do no use it. Thirdly, since the tumour dataset is stronly class-unbalanced, we hypothesized that the small classes would consist mainly of atypical data. The second and third hypotheses will be tested using the *t*-GTM variants.

For the sake of brevity, we only provide one of these illustrative visualizations in Fig. 1.

Here, two tumour groups (low grade gliomas and meningiomas) are shown. The right column of Fig. 1, where the models that include class information are located, provides some preliminary support for the first hypothesis since the class separation between both classes is better than that of the models that do not use class information, located in the left column. This can be observed in the form of a more pronounced overlapping of both classes in the left hand-side models of Fig. 1. This is reinforced by the entropy results provided later on in the paper.

The use of *t* distributions in the models represented in the bottom row yields a similar data spread to that of the standard Gaussian GTM models of the top row. This is an indication that there might be not too many clear outliers in the two classes visually represented. Therefore, the second hypothesis cannot be supported at this stage.

We now turn our attention to the third hypothesis. In (Vellido and Lisboa, 2006) it was shown that a given data instance could be characterized as an outlier if the value of

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

was sufficiently large. The histogram in Fig. 2 displays the values of O_n^* from (9) for the brain tumour dataset. We did the same for the class-*t*-GTM model and the corresponding values of O_n^* are displayed in Fig. 3.

First of all, and supporting our previous impression, not too many data could be clearly characterized as outliers according to these histograms. Somehow surprisingly, given the complex tumour typology of the dataset under study, these results do not support the third hypothesis, as most of the spectra that might be considered as outliers actually belong to the largest and best represented tumour types, such as meningiomas and glioblastomas. Interestingly, few metastases and astrocytomas are amongst the most extreme outliers.

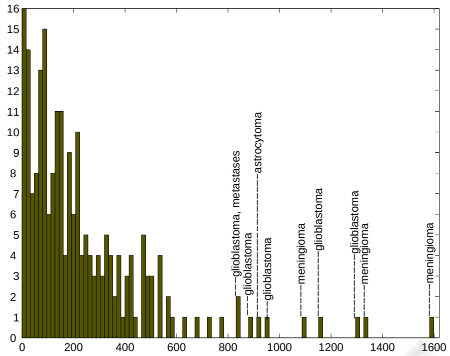

Figure 2: Histogram of the statistic (9) for the *t*-GTM model; outliers are characterized by its large values. As an example, the ten largest values are labeled.

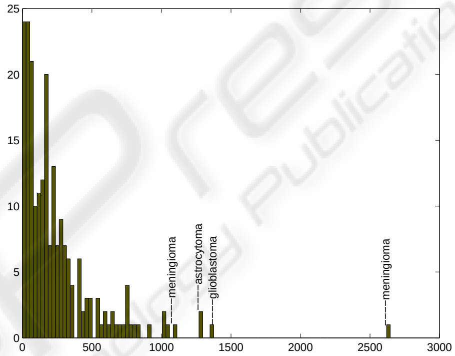

Figure 3: Histogram of (9) for class-*t*-GTM. As an example, the four largest values are labeled.

The entropy measurements quantifying the cluster-wise class separation for the brain tumour dataset are shown in Fig. 4. Two immediate conclusions can be drawn: First, all the two-stage clustering procedures based on GTM perform much better than the direct clustering of the data through K-means in terms of cluster-wise class separation. The two-stage procedure based on class-GTM also performs much better than its counterpart without class information based on the standard GTM (right hand side of Fig. 4). On the contrary, it can also be observed that the two-stage clustering based on class-*t*-GTM does not perform better than the *t*-GTM model. This is explained by the fact that the adjustment of the model provided by *t*-GTM, which is blind to class information by itself, alters the accordance between class and cluster distributions, especially in a strongly class-unbalanced dataset such as the one under analysis. This result draws the limits out of which the addition of class information is not necessarily useful in terms of cluster-wise separation. The second main conclusion to be drawn is that the random initialization in the second stage of the clustering procedure, with or without class information, does not entail any significant advantage over the proposed fixed initialization strategies across the whole range of possible final number of clusters, while being far more costly in computational terms.

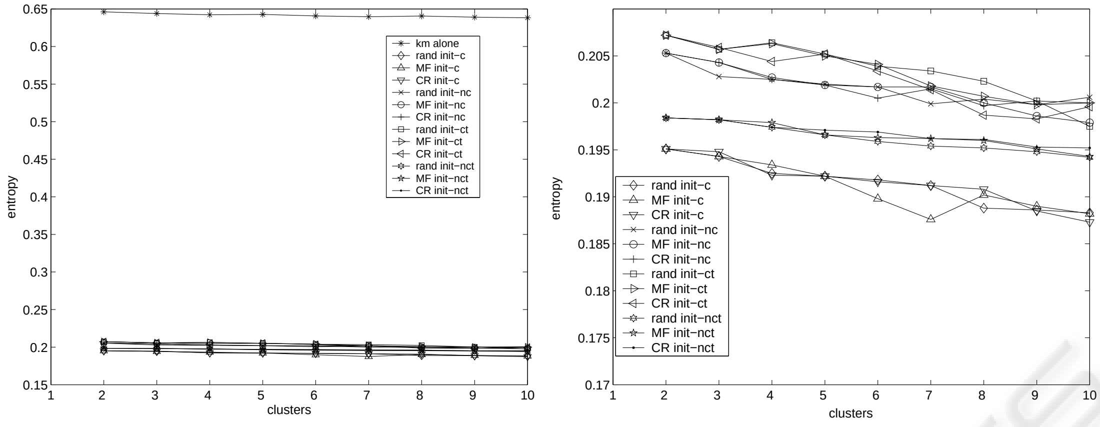

Figure 4: Entropy for the two-stage clustering of the tumour dataset, with different initializations (MF init, CR init and rand init) and K-means alone. The ‘c’ and ‘nc’ symbols refer to models that, in turn, use and not use class information. The ‘t’ in the legend means that *t*-GTM was used in the first stage. (Left): all models are shown. (Right): only the GTM, *t*-GTM and their class-enriched variants are shown.

The entropy measure in (8) quantifies the level of agreement between the clustering solutions and the class distributions. In terms of the overall cluster-wise class separation provided by the Gaussian distributions-based GTM clustering models, it has been shown that the addition of class information consistently helps. As a result, these class-enriched models would be useful in a semi-supervised setting in which new undiagnosed tumour cases were added to the database.

4 CONCLUSIONS

In this paper we have analyzed the influence exerted by the inclusion of class information in the two-stage clustering of a complex human brain tumour MRS dataset. We have also introduced two economical and principled fixed initialization procedures for the second stage of the procedure. The existence of atypical data or outliers in the human brain tumours MRS dataset under study and its influence on the clustering have also been explored.

REFERENCES

Bishop, C. M., , Svensén, M., and Williams, C. K. I. (1997). Magnification Factors for the GTM algorithm. In *Proceedings of the IEE fifth International Conference on Artificial Neural Networks*, pages 64–69.

Bishop, C. M., Svensén, M., and Williams, C. K. I. (1998). The Generative Topographic Mapping. *Neural Computation*, 10(1):215–234.

Cruz, R. and Vellido, A. (2006). On the improvement of brain tumour data clustering using class information. In *Proceedings of the 3rd European Starting AI Researcher Symposium (STAIRS’06), Riva del Garda, Italy.*

Figueiredo, M. A. T. and Jain, A. K. (2002). Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):381–396.

Kohonen, T. (1995). *Self-Organizing Maps*. Springer-Verlag, Berlin.

Sun, Y., Tiňo, P., and Nabney, I. T. (2002). Visualization of incomplete data using class information constraints. In Winkler, J. and Niranjan, M., editors, *Uncertainty in Geometric Computations*, pages 165–174. Kluwer Academic Publishers, The Netherlands.

Tate, A. R., Underwood, J., Acosta, D. M., Julià-Sapé, M., Majós, C., Moreno-Torres, A., Howe, F. A., van der Graaf, M., Lefournier, V., Murphy, M. M., Loosemore, A., Ladroue, C., Wesseling, P., Bosson, J. L., Cabañas, M. E., Simonetti, A. W., Gajewicz, W., Calvar, J., Capdevila, A., Wilkins, P. R., Bell, B. A., Rémy, C., Heerschap, A., Watson, D., Griffiths, J. R., and Arús, C. (2006). Development of a decision support system for diagnosis and grading of brain tumours using in vivo magnetic resonance single voxel spectra. *NMR in Biomedicine*, 19:411–434.

Vellido, A. (2006). Missing data imputation through GTM as a mixture of *t*-distributions. *Neural Networks*, 19(10):1624–1635.

Vellido, A. and Lisboa, P. J. G. (2006). Handling outliers in brain tumour MRS data analysis through robust topographic mapping. *Computers in Biology and Medicine*, 36(10):1049–1063.

Vesanto, J. and Alhoniemi, E. (2000). Clustering of the Self-Organizing Map. *IEEE Transactions on Neural Networks*, 11(3):586–600.