On the Initialization of Two-Stage Clustering with Class-GTM

Raúl Cruz-Barbosa[1,2] and Alfredo Vellido[1]

[1] Universitat Politècnica de Catalunya, Jordi Girona, 08034, Barcelona, Spain
{rcruz,avellido}@lsi.upc.edu
www.lsi.upc.edu/dept/investigacion/sectia/soco
[2] Universidad Tecnológica de la Mixteca, Car. Acatlima km. 2.5, 69000, Huajuapan, Oaxaca, México

Abstract. Generative Topographic Mapping is a probabilistic model for data clustering and visualization. It maps points, considered as prototype representatives of data clusters, from a low dimensional latent space onto the observed data space. In semi-supervised settings, class information can be added resulting in a model variation called class-GTM. The number of class-GTM latent points used is usually large for visualization purposes and does not necessarily reflect the class structure of the data. It is therefore convenient to group the clusters further in a two-stage procedure. In this paper, class-GTM is first used to obtain the basic cluster prototypes. Two novel methods are proposed to use this information as prior knowledge for the K-means-based second stage. We evaluate, using an entropy measure, whether these methods retain the class separability capabilities of class-GTM in the two-stage process, and whether the two-stage procedure improves on the direct clustering of the data using K-means.

1 Introduction

Amongst density-based methods, Finite Mixture Models have established themselves as a flexible and robust tool for multivariate data clustering [1]. In many practical data analysis scenarios, though, the available knowledge concerning the cluster structure of the data may be quite limited. In these cases, data exploration techniques are valuable tools and, amongst them, multivariate data visualization can be of great help by providing the analyst with intuitive cues about data structural patterns. In order to endow Finite Mixture Models with data visualization capabilities, certain constraints must be enforced. One alternative is forcing the model components to be centred in a low-dimensional manifold embedded into the usually high-dimensional observed data space. Such approach is the basis for the definition of Generative Topographic Mapping (GTM) [2], a flexible manifold learning model for simultaneous data clustering and visualization whose probabilistic nature makes possible to extend it to perform tasks such as missing data imputation [3], robust handling of outliers [4], and unsupervised feature selection [5], amongst others.

D. Borrajo, L. Castillo, and J.M. Corchado (Eds.): CAEPIA 2007, LNAI 4788, pp. 50–59, 2007.

Finite Mixture Models can also be used beyond unsupervised learning in order to account for class-related information in supervised or semi-supervised settings [6]. Class information can be integrated as part of the GTM training to enrich the cluster structure definition provided by the model [7,8]. The resulting class-GTM model is the basis of this paper.

GTM in general and class-GTM in particular 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 distributions of the data. For that reason, a two-stage clustering procedure may be useful in this scenario [9]. Class-GTM 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, also known as prototypes, can be further clustered in the second stage. For that role, the well-known K-means algorithm is used in this study. The issue remains of how we should initialize K-means in the second clustering stage. Random initialization, with the subsequent choice of the best solution, was the method selected in [9]. This approach, though, does not make use of the prior knowledge generated in the first stage of the procedure. Here, we propose two different ways of introducing such prior knowledge in the initialization of the second stage K-means, without compromising the final clusterwise class separation capabilities of the model. This fixed initialization procedures allow significant computational savings.

The outline of the remaining of the paper is as follows: In section 2, we summarily introduce the GTM and its class-GTM variant, 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 and directions for future research.

2 Two-Stage Clustering

The two-stage clustering procedure outlined in the introduction is described in this section. The first stage model, namely class-GTM, is introduced first. This is followed by the details of different initialization strategies for the second stage. We propose two novel second stage fixed initialization strategies that take advantage of the prior knowledge obtained in the first stage.

2.1 The Class-GTM Model

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 is a set of M basis functions $\phi(\mathbf{u}) = (\phi_\mathbf{1}(\mathbf{u}), ..., \phi_\mathbf{M}(\mathbf{u}))$. For continuous data of dimension D, spherically symmetric Gaussians

$$\phi_m(\mathbf{u}) = \exp\left\{-1/2\sigma^2 \|\mathbf{u} - \mu_m\|^2\right\} \tag{2}$$

are an obvious choice of basis function, with centres μ_m and common width σ; $\mathbf{W}$ 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:

$$L(\mathbf{W}, \beta|\mathbf{X}) = \sum_{n=1}^{N} \ln \left\{ \frac{1}{K} \sum_{k=1}^{K} \left(\frac{\beta}{2\pi}\right)^{D/2} \exp\left\{-\beta/2 \|\mathbf{y}_k - \mathbf{x}_n\|^2\right\} \right\} \tag{3}$$

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 variance, which accounts for the fact that data points might not strictly lie on the low dimensional embedded manifold generated by the GTM. 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 and therefore inherits most of its properties. The main goal of this extension is to improve class separability in the clustering results of GTM. For this purpose, we assume that the clustering model accounted for the available class information. This can be achieved by modelling the joint density $p(C, \mathbf{X})$, instead of $p(\mathbf{X})$, for a given set of classes $\{T_i\}$. For the Gaussian version of the GTM model [7,8], such approach entails the calculation of the posterior probability of a cluster representative $\mathbf{u}_k$ given the data point $\mathbf{x}_n$ and its corresponding class label c_n, or class-conditional *responsibility* $\hat{z}^c_{kn} = p(\mathbf{u}_k|\mathbf{x}_n, c_n)$, as part of the E step of the EM algorithm. It can be calculated as:

$$\hat{z}^c_{kn} = \frac{p(\mathbf{x}_n, c_n|\mathbf{u}_k)}{\sum_{k'=1}^{K} p(\mathbf{x}_n, c_n|\mathbf{u}_{k'})} = \frac{p(\mathbf{x}_n|\mathbf{u}_k)p(c_n|\mathbf{u}_k)}{\sum_{k'=1}^{K} p(\mathbf{x}_n|\mathbf{u}_{k'})p(c_n|\mathbf{u}_{k'})} = \frac{p(\mathbf{x}_n|\mathbf{u}_k)p(\mathbf{u}_k|c_n)}{\sum_{k'=1}^{K} p(\mathbf{x}_n|\mathbf{u}_{k'})p(\mathbf{u}_{k'}|c_n)}, \tag{4}$$

and, being T_i each class,

$$p(\mathbf{u}_k|T_i) = \frac{\sum_{n;c_n=T_i} p(\mathbf{x}_n|\mathbf{u}_k) / \sum_n p(\mathbf{x}_n|\mathbf{u}_k)}{\sum_{k'} \sum_{n;c_n=T_i} p(\mathbf{x}_n|\mathbf{u}_{k'}) / \sum_n p(\mathbf{x}_n|\mathbf{u}_{k'})} \ . \tag{5}$$

Equation (4) differs from the standard responsibility $\hat{z}_{kn}$ of GTM in that, instead of imposing a fixed prior $p(\mathbf{u}_k) = 1/K$ on latent space, we consider a

class-conditional prior $p(\mathbf{u}_k|T_i)$. Once the class-conditional responsibility is calculated, the rest of the model's parameters are estimated following the standard EM procedure.

2.2 Two-Stage Clustering Based on GTM

In the first stage of the proposed two-stage clustering procedure, a class-GTM is trained to obtain the representative prototypes (detailed clustering) of the observed dataset $\mathbf{X}$. As mentioned in the introduction, the number of prototype vectors is usually chosen to be large for visualization purposes, and does not necessarily reflect the global cluster and class structure of the data. In this study, the resulting prototypes $\mathbf{y}_k$ of the class-GTM are further clustered using the K-means algorithm. In a similar two-stage procedure to the one described in [9], 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

$$E = \sum_{k=1}^{C} \sum_{\mathbf{x} \in G_k} \|\mathbf{x} - \mu_k\|^2, \tag{6}$$

where C is the final number of clusters in the second stage and μ_k is the centre of cluster G_k. 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 take advantage of the prior knowledge obtained by class-GTM in the first stage are proposed. They are based on two features of the model, namely: the Magnification Factors (MF) and the Cumulative Responsibility (CR). The Magnification Factors measure the level of distorsion of the mapping from the latent to the data spaces. Areas of low data concentration correspond to high distorsions of the mapping (i.e., 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:

$$\frac{dA'}{dA} = \det{}^{1/2}\left(\psi^{\mathbf{T}}\mathbf{W}^{\mathbf{T}}\mathbf{W}\psi\right), \tag{7}$$

where ψ has elements $\psi_{ji} = \partial\phi_j/\partial u^i$ [10]. 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.

As its name suggests, the CR is the sum of responsibilities over all data points in $\mathbf{X}$ for each cluster k:

$$CR_k = \sum_{n=1}^{N} \hat{z}_{kn}^{c} \quad . \tag{8}$$

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 class-GTM prototypes corresponding to the C non-contiguous latent points with highest CR. That is, the second stage algorithm is meant to start from those cluster prototypes that are found to be most responsible for the generation of the observed data.

3 Experiments

In this section, we first describe the experimental design and settings. This is followed by a presentation and discussion of the corresponding results.

3.1 Experimental Design and Settings

The class-GTM model was implemented in MATLAB®. For the experiments reported next, the adaptive matrix $\mathbf{W}$ was initialized, following a procedure described in [2], as to minimize the difference between the prototype vectors y_k and the vectors that would be generated in data space by a partial PCA, $m_k = V_2 u_k$, where the columns of matrix V_2 are the two principal eigenvectors (given that the latent space considered here is 2-dimensional). Correspondingly, the inverse variance β was initialised to be the inverse of the 3^{rd} PCA eigenvalue. This ensures the replicability of the results. The value of parameter σ, describing the common width of the basis functions, was set to 1. The grid of latent points u_k was fixed to a square 13x13 layout for the *ecoli* dataset and to a 20x20 layout for the *oil flow* dataset. Both datasets are summarily described in section 3.2. The corresponding grid of basis functions ϕ was equally fixed to a 5x5 square layout for both datasets.

The goals of these experiments are twofold. First, we aim to assess whether a two-stage clustering procedure, where the first stage involves class-GTM and the second stage involves K-means, improves on the class separation capabilities of the straight clustering of the data using the K-means algorithm alone. Secondly, we aim to test whether the second stage initialization procedures based on the Magnification Factors and the Cumulative Responsibility of the class-GTM, described in section 2.2, retain the class separability capabilities of the two-stage clustering procedure in which K-means is randomly initialized. If this is the case, a fixed second stage initialization strategy should entail a substantial reduction of computational time compared to a random second stage initialization requiring a large number (100 in the reported experiments and also in [9]) of algorithm runs.

Beyond the visual exploration that could be provided by class-GTM, the second stage clustering results should be explicitly quantified in terms of class separability. For that purpose, the following entropy-like measure is proposed:

$$E_{C_k}(\{T_i\}) = -\sum_{\{C_k\}} P(C_k) \sum_{\{T_i\}} P(T_i|C_k) \ln P(T_i|C_k) = -\sum_{k=1}^{K} \frac{N_k}{N} \sum_{i=1}^{|\{T_i\}|} p_{ki} \ln p_{ki} \ . \tag{9}$$

Sums are performed over the set of classes $\{T_i\}$ and the class-GTM clusters $\{C_k\}$; N is the total number of prototypes; N_k is the number of prototypes assigned to the k^{th} cluster; $p_{ki} = \frac{n_{ki}}{N_k}$, where n_{ki} is the number of prototypes from class i assigned to cluster k; and, finally, $|\{T_i\}|$ is the cardinality of the set of classes. The minimum possible entropy value is 0, which 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. This is a time-honoured matter of debate, which goes beyond the scope of this paper, and many cluster validity indices have been defined over the years. In this paper we use the widely known Davies-Bouldin (DB) index [11,9] to provide us with some indication of what the adequate number of final clusters might be. According to the DB index, the best clustering minimizes

$$\frac{1}{C} \sum_{k=1}^{C} \max_{l \neq k} \left\{ \frac{S_c(G_k) + S_c(G_l)}{d_{ce}(G_k, G_l)} \right\}, \tag{10}$$

where C is the number of clusters; S_c is a within-cluster distance named centroid distance and is calculated as $S_c = \frac{\sum_{\mathbf{y}_i \in G_k} \|\mathbf{y}_i - \mu_k\|}{N_k}$, N_k is the number of samples in cluster G_k; and d_{ce} is a between-clusters distance named centroid linkage defined as $d_{ce}(G_k, G_l) = \|\mu_k - \mu_l\|$.

3.2 Results and Discussion

In the first stage of the two-stage clustering procedure, class-GTM was trained to model two datasets taken, in turn, from the UCI and the Pattern Recognition and Machine Learning book[1] repositories: *ecoli* and *oil flow*. The resulting prototypes $\mathbf{y}_k$ were then clustered in the second stage using the K-means algorithm. This last stage was performed in three different ways, as described in section 2.
In the first one, K-means was randomly initialized 100 times, selecting the results corresponding to the minimum of the error function in (6). In the second, we used the Magnification Factors of class-GTM as prior knowledge for the initialization of K-means. In the third, Cumulative Responsibility was used as prior knowledge. In all cases, K-means was forced to yield a given number of final clusters, from 2 up to 13. The DB index and the final entropy were calculated for all the above procedures and numbers of clusters.

The DB index results for the experiments with *ecoli*, including the direct clustering of the data with K-means alone, are reported in Fig. 1. *Ecoli* consists of 336 7-dimensional points belonging to 8 classes representing protein location

[1] http://research.microsoft.com/∼cmbishop/PRML/webdatasets/datasets.htm

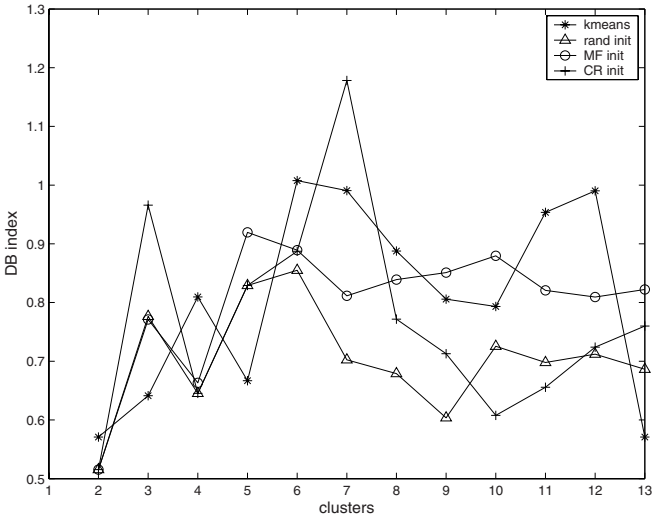

Fig. 1. DB index for the clustering of *ecoli* using two-stage clustering with different initializations (based on Magnification Factors (MF init), Cumulative Responsibility (CR init) and random (rand init)), and K-means alone

sites, 3 of which are very small, i.e., the data set is strongly class-unbalanced. It is therefore unsurprising that the results in Fig. 1 do not provide a clear pattern. They nevertheless suggest that no more of 4 clusters (for two-stage clustering) or 5 (for direct K-means) represent an adequate solution. In fact, there are only 4 main groups in *ecoli*, namely: cytoplasm, periplasm, inner membrane and outer membrane. Some relatively good solutions are also suggested for 8 or 9 clusters using the two-stage procedure.

The entropy results for *ecoli* are shown in Fig. 2. Two immediate conclusions can be drawn: First, all the two-stage clustering procedures based on class-GTM perform much better than direct K-means clustering in terms of class separation in the resulting clusters. Second, random initialization in the second stage of the clustering procedure 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.

The DB index results for the experiments with *oil flow*, also including the direct clustering of the data with K-means, are reported in Fig. 3. *Oil flow*, firstly used in [12], simulate non-intrusive measurements by gamma densitometry from a pipeline transporting a mixture of gas, oil, and water. It consists of 1000 points described by 12 attributes. Three types of flow configuration are used as class information labels. The results in Fig. 3 do not indicate any clear number of clusters when data are grouped directly by K-means without any class information. Instead, for the two-stage procedure based on class-GTM there is no indication that more than 4 clusters would provide any substantial improvement.

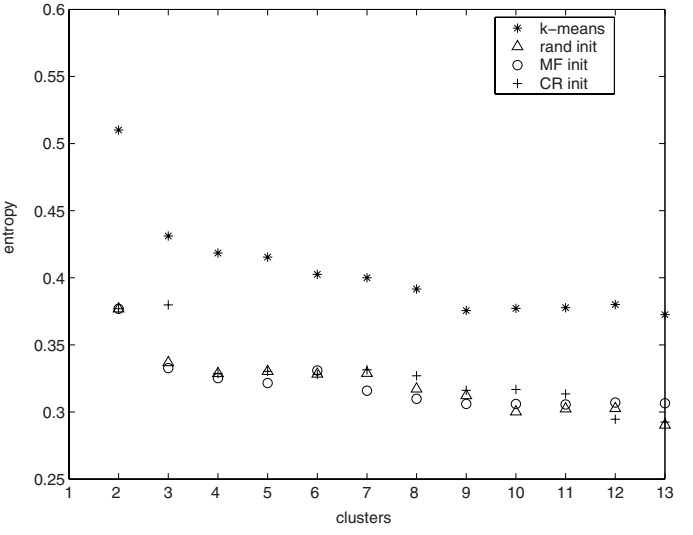

Fig. 2. Entropy measurements for two stage and K-means alone clusterings of *ecoli*. Legend as in Fig. 1.

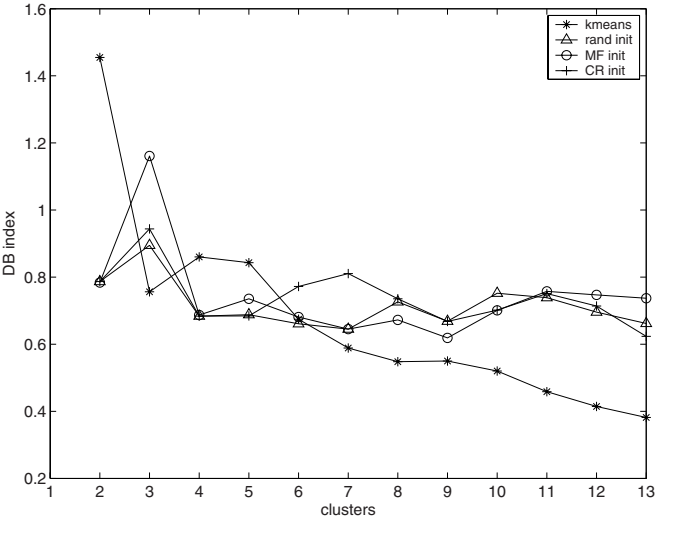

Fig. 3. DB index for the clustering of *oil flow* using two-stage clustering with different initializations and K-means alone. Legend as in Fig. 1.

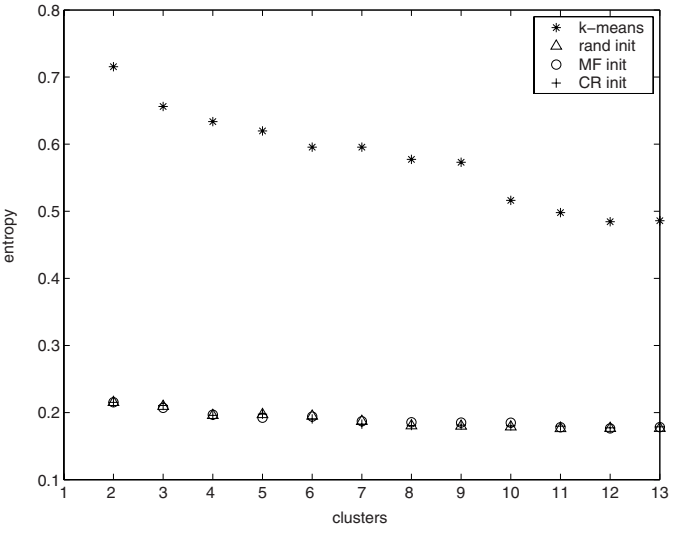

Fig. 4. Entropy measurements for two stage and K-means alone clusterings of *oil flow*. Legend as in Fig. 1.

The entropy results for *oil flow* are shown in Fig. 4 and they are fully consistent with the results for *ecoli*. Again, the two-stage clustering procedures based on class-GTM perform much better than direct K-means clustering in terms of class separation, and the two-stage random and fixed initialization strategies yield almost identical results, with the former being computationally more costly.

4 Conclusion

In this paper we have analysed different strategies of initialization for a two-stage multivariate data clustering procedure. The first stage is based on the manifold learning class-GTM model, which, besides clustering, also provides data and clusters visualization on a low-dimensional space. The second stage is based on the well-known K-means algorithm, which was initialized either multiple times randomly or, making use of the prior knowledge provided by class-GTM in the first stage, in a fixed manner using a novel procedure based on its Magnification Factors and Cumulative Responsibility. Several experiments have shown that the two-stage random and fixed initializations yield almost identical results in terms of clusterwise class separation, with the former being computationally more costly. It has also been shown that the two-stage clustering procedures based on class-GTM perform much better than direct K-means clustering of the data in terms of this clusterwise class separation.

Future research should expand the reach of the analyses to assess whether the class information-enriched class-GTM model performs better than the standard GTM as first stage clustering method. Different cluster validity indices could

also be used in order to obtain further guidance on the choice of an appropriate final number of clusters.

Acknowledgements. Alfredo Vellido is a researcher within the Ramón y Cajal program of the Spanish MEC and acknowledges funding from the MEC I+D project TIN2006-08114. Raúl Cruz-Barbosa acknowledges SEP-SESIC (PROMEP program) of México for his PhD grant.

References

1. Figueiredo, M.A.T., Jain, A.K.: Unsupervised learning of finite mixture models. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(3), 381–396 (2002)
2. Bishop, C.M., Svensén, M., Williams, C.K.I.: The Generative Topographic Mapping. Neural Computation 10(1), 215–234 (1998)
3. Vellido, A.: Missing data imputation through GTM as a mixture of t-distributions. Neural Networks 19(10), 1624–1635 (2006)
4. Vellido, A., Lisboa, P.J.G.: Handling outliers in brain tumour MRS data analysis through robust topographic mapping. Computers in Biology and Medicine 36(10), 1049–1063 (2006)
5. Vellido, A., Lisboa, P.J.G., Vicente, D.: Robust analysis of MRS brain tumour data using t-GTM. Neurocomputing 69(7-9), 754–768 (2006)
6. Hastie, T., Tibshirani, R.: Discriminant analysis by Gaussian mixtures. Journal of the Royal Statistical Society (B) 58, 155–176 (1996)
7. Cruz, R., Vellido, A.: 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 (2006)
8. Sun, Y., Tiňo, P., Nabney, I.T.: Visualization of incomplete data using class information constraints. In: Winkler, J., Niranjan, M. (eds.) Uncertainty in Geometric Computations, pp. 165–174. Kluwer Academic Publishers, The Netherlands (2002)
9. Vesanto, J., Alhoniemi, E.: Clustering of the Self-Organizing Map. IEEE Transactions on Neural Networks (2000)
10. Bishop, C.M., Svensén, M., Williams, C.K.I.: Magnification Factors for the GTM algorithm. In: Proceedings of the IEE fifth International Conference on Artificial Neural Networks, pp. 64–69 (1997)
11. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Trans. on Pattern Analysis and Machine Intelligence 1(2), 224–227 (1979)
12. Bishop, C.M., James, G.D.: Analysis of multiphase flows using dual-energy gamma densitometry and neural networks. Nuclear Instruments and Methods in Physics Research A327, 580–593 (1993)