Proceedings of the 5th WSEAS Int. Conf. on APPLIED INFORMATICS and COMMUNICATIONS, Malta, September 15-17, 2005 (pp270-275)

Faster clustering of complex data with the Generalised Harmonic Topographic Mapping (G-HaToM)

MARIAN PEÑA AND COLIN FYFE
Applied Computational Intelligence Research Unit,
The University of Paisley,
Paisley, PA1 2BE
SCOTLAND.

Abstract In this paper we explain the Generalised Harmonic topographic Map (G-HaToM), an extension of the Harmonic Topographic map [3] and [4]. This algorithm extends the mapping from data space to latent space using the p^{th} power of the L^2 distance, where the second power is the former version of the topographic mapping, HaToM. This generalization allows the mapping of more difficult data, reducing at the same time the computational cost of the mapping for the data already clustered by the original HaToM.

Key-words: Smooth manifold identification, Clustering, Topographic maps, Boosting.

1 Introduction

The Harmonic Topographic Map (HaToM) was developed as a clustering alternative to the ToPoE [1], which is also based on the GTM. The HaToM has the same structure as the GTM, with a number of latent points that are mapped to a feature space by M Gaussian functions, and then into the data space by a matrix W. Each latent point, indexed by k is mapped, through a set of M basis functions, $\Phi_1(), \Phi_2(), \cdots, \Phi_M()$ to a centre in data space, $\mathbf{m}_k = \Phi_k W$. But the similarity ends there because the objective function is not the GTM one, neither is it optimised with the Expectation-Maximization (EM) algorithm. Instead, the HaToM uses the well proved clustering abilities of the K-means algorithm, improved by using harmonic means to make it insensitive to initialisation ([8]).

In this paper we extend the algorithm using the p^{th} power of the L^2 distance that gives it a boosting-like property that helps the algorithm to get faster clustering, specially for difficult data.

In the rest of the paper we develop the algorithm, reviewing first the properties of the Model-driven Harmonic mapping (M-HaToM) [3], and generalizing then with the p^{th} power of the L^2 distances. Finally we present a few experiments comparing the M-HaToM, which is equivalent to the second power of the L^2 distance, with higher powers of the G-HaToM.

2 Model-driven HaToM

In [3] and [4] we developed two versions of the Harmonic mapping. In this paper, we compared the generalization G-HaToM with the Model-driven algorithm (M-HaToM).

The M-HaToM algorithm is

1. Initialise K to 2. Initialise the W weights randomly and spread the centres of the M basis functions uniformly in latent space.

2. Initialise the K latent points uniformly in latent space. Set count=0.

3. Calculate the projection of the latent points to data space. This gives the K centres, $\mathbf{m}_k = \phi_k^T W$.

4. For every data point, $\mathbf{x}_i$, calculate $d_{i,k} = ||\mathbf{x}_i - \mathbf{m}_k||$.

5. Recalculate centres using

$$\mathbf{m}_k = \frac{\sum_{i=1}^{N} \frac{1}{d_{i,k}^4 (\sum_{l=1}^{K} \frac{1}{d_{i,l}^2})^2} \mathbf{x}_i}{\sum_{i=1}^{N} \frac{1}{d_{i,k}^4 (\sum_{l=1}^{K} \frac{1}{d_{i,l}^2})^2}} \quad (1)$$

6. Recalculate W using

$$W = \begin{cases} (\Phi^T\Phi + \delta I)^{-1}\Phi^T\mathbf{\Xi} & \text{if } K < M \\ (\Phi^T\Phi)^{-1}\Phi^T\mathbf{\Xi} & \text{if } K \geq M \end{cases}$$

where $\mathbf{\Xi}$ is the matrix containing the K centres, I is identity matrix and δ is a small constant, necessary because initially $K < M$ and so the matrix $\Phi^T\Phi$ is singular.

7. If count<MAXCOUNT, count= count +1 and return to 3

8. If $K < K_{max}$, $K = K + 1$ and return to 2.

If we wish to use the mapping for visualisation, we must map data points into latent space. To do this, we define the responsibility that the k^{th} latent point has for the i^{th} data point as

$$r_{ik} = \frac{\exp(-\gamma d_{i,k})}{\sum_{l=1}^{K} \exp(-\gamma d_{i,l})} \quad (2)$$

and the new data point is placed at y_i where

$$y_i = \sum_{k=1}^{K} r_{i,k} t_k \quad (3)$$

where t_k is the position of the k^{th} latent point in latent space. γ is known as the width of the responsibilities.

[1] http://www.stats.ox.ac.uk/pub/PRNN/crabs.dat

3 Generalised Harmonic Topographic Map (G-HaToM)

The Generalised version of the algorithm (G-HaToM) includes the p^{th} power of the L^2 distance which have a "Dynamic weighting function" [7] that determines how data points participates in the next iteration to calculate the new centers $\mathbf{m}_k$. The weight is bigger for the data points further away from the centres, so that their participation is boosted in the next iteration. This makes the algorithm insensitive to initialisation and also prevents one cluster from taking more than one centre.

The only change from the M-HaToM is in the recalculation of the centres, which in this case is:

$$\mathbf{m}_k = \frac{\sum_{i=1}^{N} \frac{1}{d_{i,k}^p (\sum_{l=1}^{K} \frac{1}{d_{i,l}^2})^{p-2}} \mathbf{x}_i}{\sum_{i=1}^{N} \frac{1}{d_{i,k}^p (\sum_{l=1}^{K} \frac{1}{d_{i,l}^2})^{p-2}}} \quad (4)$$

where p is the p^{th} power of the L^2 distance allocated.

4 Simulations:

The p^{th} power of the L^2 distances are able to separate better into clusters high dimensional and also more complex data, such as the crabs or the oil data (see below). Also, the boosting-like weighting allows the algorithm to get faster to the clustering as we will see comparing both HaToM with the algae data.

4.1 Crabs Data

This is a 5 dimensional dataset[1] on the morphology of rock crabs of genus Leptograpsus, with 50 specimens of each sex of each of two colour forms, blue and orange. This data is used in the Generative Topographic Map(GTM) PhD by Svensén [5] to show the projection into latent space of the four clusters

with the GTM. We illustrate the results of the G-HaToM algorithm in Figure 1, using the L^2 distance to the third power, 40 centers, 20*20 latent points and 20 iterations, over non normalised data (unlike the GTM which proved to be better with normalised crab data). The projection keeps together the two female clusters on the low part of the figure, while the male clusters are at the top; only the blue form sexes stay closer.

4.2 Oil Data

The oil flow dataset[2] consists of 1000 points classified into three flow configurations. This is synthetic data modelling non-intrusive measurements on a pipe-line transporting a mixture of oil, water and gas. The flow in the pipe takes one out of three possible configurations: horizontally stratified, nested annular or homogeneous mixture flow. The data lives in a 12-dimensional measurement space, but for each configuration, there is only two degrees of freedom: the fraction of water and the fraction of oil. (The fraction of gas is redundant, since the three fractions must sum to one.) Hence, the data lives on a number of 'sheets' which locally are approximately 2-dimensional. The data is 12 dimensional and therefore more suitable for the purpose of showing the capabilities of an algorithm to classify complex data sets where a single two-dimensional visualization plot may not be enough. This data is used to check the hierarchical GTM in [6].

In this case again the p^{th} power of the L^2 distance was better (compared to the previous HaToM) to separate the clusters and Figure 2 shows the projection onto a 2 dimensional map with 60 by 60 latent points, 40 iterations and 20 centre points. The L^2 distance was to the power of 5 . The advantage in comparison with the hierarchical GTM is the simplicity and the computational cost.

4.3 Algae Data

This is a set of 118 samples from a scientific study of various forms of algae some of which have been manually identified. Each sample is recorded as an 18 dimensional vector representing the magnitudes of various pigments. 72 samples have been identified as belonging to a specific class of algae which are labeled from 1 to 9. 46 samples have yet to be classified and these are labeled 0.

The M-HaToM gives a very good clustering of this data as we showed in [3] and [4]. To illusrate the improvement in computational cost with the G-HaToM, we reduce the number of latent points to the minimum: G-HaToM is able to cluster this data with only 4*4 latent points and p=3 in 2.41 seconds, as shown in Figure 3, while the M-HaToM needs at least 5*5 latent points and requires 4.34 seconds to do so (see Figure 4).

5 Conclusion

We introduced an extension of the Harmonic Topographic map that has a boosting property, allowing the data to get a faster clustering, helping the data points further away to have a bigger weight in the next iteration. Future work will include a most extensive study of the generalisations of both, model and data driven HaToM (M-HaToM and D-HaToM) with more complex data. We will also analyze the G-HaToM with different initialisations to prove that the algorithm is not sensitive to them.

[2] http://www.ncrg.aston.ac.uk/GTM/

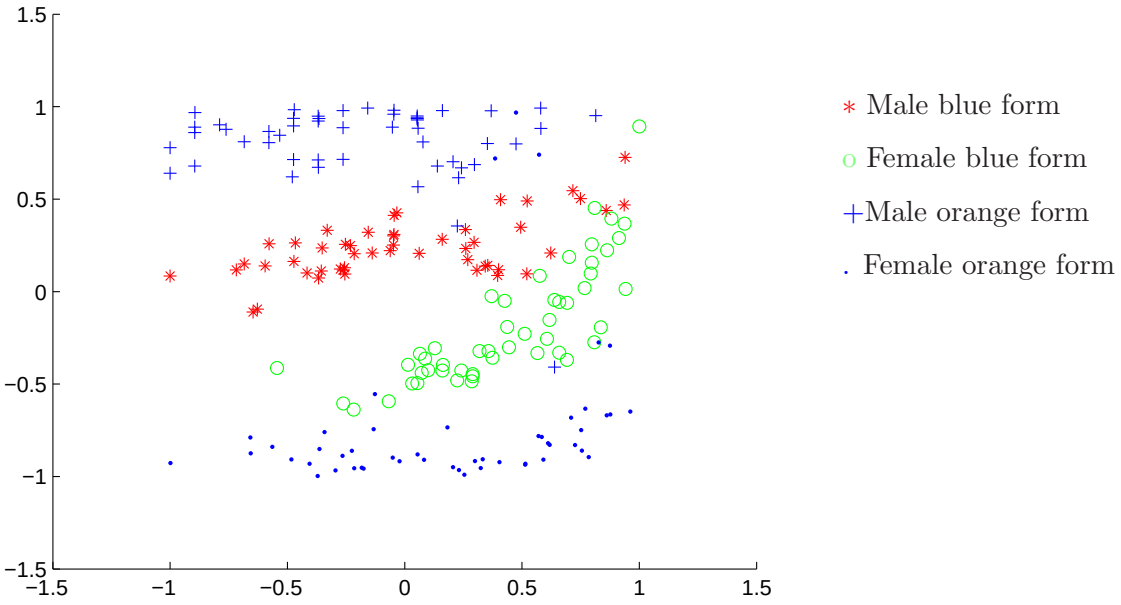

Figure 1: G-HaToM projection of the two species of crabs with equal proportion of both sexes: power=3

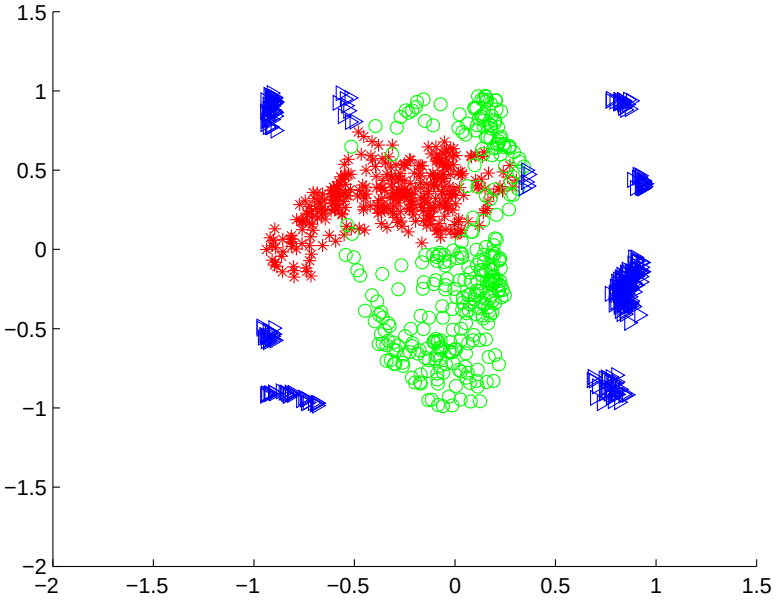

Figure 2: G-HaTom projection of the oil data: power=5

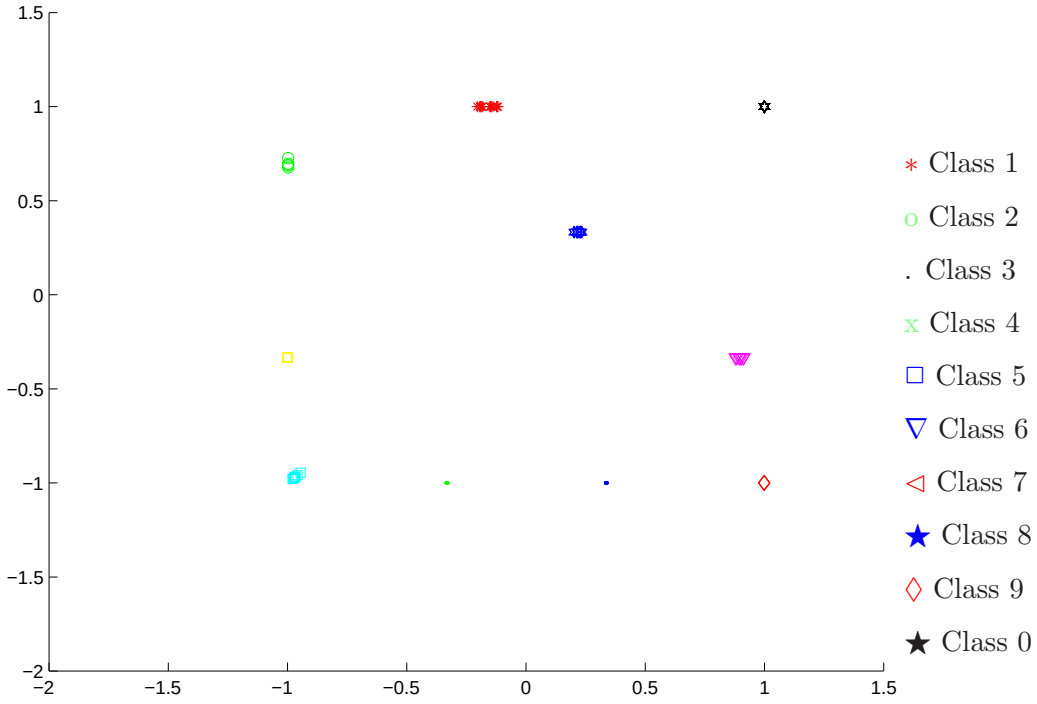

Figure 3: G-HaTom projection of the algae data: power=3 and 4*4 latent points.

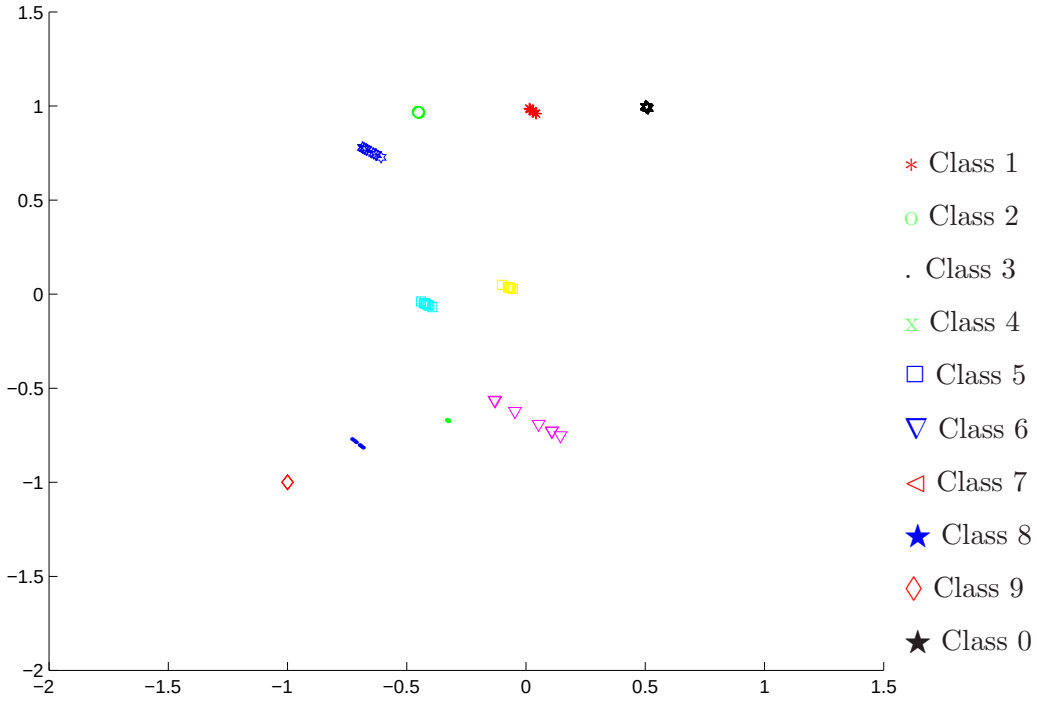

Figure 4: M-HaTom projection of the algae data: 5*5 latent points.

References

[1] C. Fyfe. The topographic product of experts. In *International Conference on Artificial Neural Networks, ICANN2005*, 2005.

[2] M. Peña and C. Fyfe. The harmonic topographic map. In *The Irish conference on Artificial Intelligence and Cognitive Science, AICS05*, 2005.

[3] M. Peña and C. Fyfe. Model- and data-driven harmonic topographic maps. *WSEAS TRANSACTIONS ON COMPUTERS*, 2005.

[4] M. Peña and C. Fyfe. Tight clusters and smooth manifolds with the harmonic topographic map. In *5th WSEAS International Conference on SIMULATION, MODELING and OPTIMIZATION, WSEAS SMO '05*, 2005.

[5] M. Svensén. *GTM: The Generative Topographic Mapping.* PhD thesis, Aston University, Birmingham, UK, 1998.

[6] P. Tino and I. Nabney. Hierarchical GTM: constructing localized non-linear projection manifolds in a principled way. *(IEEE) Transactions on Pattern Analysis and Machine Intelligence, in print.*, 2001.

[7] B. Zhang. Generalized k-harmonic means – boosting in unsupervised learning. Technical report, HP Laboratories, Palo Alto, October 2000.

[8] B. Zhang, M. Hsu, and U. Dayal. K-harmonic means - a data clustering algorithm. Technical report, HP Laboratories, Palo Alto, October 1999.