ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 477-482

Using Andrews Curves for Clustering and Sub-clustering Self-Organizing Maps

César García-Osorio[1], Jesús Maudes[1], Colin Fyfe[2]

[1] Departamento de Ingeniería Civil,
Universidad de Burgos, Spain,
{cgosorio|jmaudes}@ubu.es.

[2]Applied Computational Intelligence Research Unit,
The University of Paisley, Scotland,
colin.fyfe@paisley.ac.uk.

Abstract. The use of self-organizing maps to analyze data often depends on finding effective methods to visualize the SOM's structure. In this paper we propose a new way to perform that visualization using a variant of Andrews' Curves. Also we show that the interaction between these two methods allows us to find sub-clusters within identified clusters.

1 Introduction

A Self Organising Map (SOM) is a well known method for clustering data [6]: it clusters in such a way that data point which are alike in some way tend to be matched to neurons which are alike in some way and only such data points are so matched. The SOM imposes a strict structure on its output neurons: they will often lie on a line or on a two-dimensional grid. However, the important thing about the trained SOM is that its centres or model vectors lie in the data space which may be very high dimensional. Now if we take a black box approach we may simply take the clusters which the SOM has found. However if we are interested in exploratory data analysis such as data mining, we often wish to try to understand the shape of the data structure in its high dimensional space. This paper deals with this problem.

After training the SOM, the identification of clusters is not very difficult if we use an appropriate way to visualize the structure that the SOM has acquired. There are several ways to perform this visualization; in [7] a summary of such methods is presented. The visualization of the SOM enables us to perform an interactive type of data mining that has proven to be useful in exploratory data analysis [5]

In [4] we have investigated the use of Andrews' Curves as a way to get a picture of the structure of the feature space defined by a kernel matrix. In this paper we apply the same idea and we analyze the use of Andrews' Curves as a visualization tool for the SOM's structure. We can use the Andrews' Curves to visualize the model vectors or centres of the SOM, that in general are multidimensional and so difficult to comprehend by inspection. We can use the Andrews' Curves to identify subclusters within the data set projected to a neuron of the SOM. Finally, we can use the Andrews' Curves to cluster the SOM, that is to discover groups of model vectors that form clusters; and to construct dynamic projections of the SOM grid.

2 Andrews Curves

Andrews [1] described his curves in 1972, early on in the computing era; it is an interesting observation that he thought it necessary to counsel "an output device with relatively high precision ... is required". Current standard PC hardware and software are quite sufficient for the purposes. The method is another way to attempt to visualise and hence to find structure in high dimensional data. Each data point $\mathbf{x} = \{x_1, x_2, ..., x_n\}$ defines a function

$$f_{\mathbf{x}}(t) = x_1/\sqrt{2} + x_2 \sin(t) + x_3 \cos(t) + x_4 \sin(2t) + x_5 \cos(2t) + ... \qquad (1)$$

and this function is then plotted for $-\pi < t < \pi$. Thus each data point may be viewed as a line between $-\pi$ and π. The interesting thing is that this function representation preserves distances, that is close points will appear as close curves and distant points as distant curves. If there is structure in the data, it may be visible in the Andrews' Curves of the data. The curves themselves can be considered as the projections of each point into the vector

$$\mathbf{w} \quad = \quad (1/\sqrt{2}, \sin(t), \cos(t), \sin(2t), \cos(2t), ...) \qquad (2)$$

Wegman and Solka [9] discuss the benefits of using a slightly different projection, namely that onto

$$\begin{aligned} \mathbf{w}_1 &= \sqrt{2/n}(\sin(\lambda_1 t), \cos(\lambda_1 t), ..., \sin(\lambda_{\frac{n}{2}} t), \cos(\lambda_{\frac{n}{2}} t)) \\ \mathbf{w}_2 &= \sqrt{2/n}(\cos(\lambda_1 t), -\sin(\lambda_1 t), ..., \cos(\lambda_{\frac{n}{2}} t), -\sin(\lambda_{\frac{n}{2}} t)) \end{aligned}$$

with the λ_j linearly independent over the rationals, now the curves are not periodic and it makes sense to draw them for t outside the interval $(-\pi, \pi)$. Clearly, $(\mathbf{w}_1, \mathbf{w}_2)$ form a set of 2 orthonormal basis vectors. If we define $y_1 = \mathbf{w}_1^T \mathbf{x}$, $y_2 = \mathbf{w}_2^T \mathbf{x}$ then we have a two dimensional display on which to project $\mathbf{x}$ so that we can look for structure by eye. Visually from this projection, we can identify clusters of points which are nearby and whose trajectories as we change t (i.e. as we move along the Andrews' Curves) keep close together. When we use these curves in this way we obtain a two dimensional grand tour [2] of the data (to be more precise we obtain what Wegman call a pseudo grand tour).

We have proposed in [3] an extension of this idea that make use of three different orthogonal vectors so that now $\mathbf{w}_i = f(t, s)$ such as

$$\begin{aligned} y_1 = \mathbf{w}_1^T \mathbf{x} &\propto x_1 \cos(\lambda_1 t)\cos(\mu_1 s) + x_2 \cos(\lambda_1 t)\sin(\mu_1 s) + x_3 \sin(\lambda_1 t) + ... \\ y_2 = \mathbf{w}_2^T \mathbf{x} &\propto x_1 \sin(\lambda_1 t)\cos(\mu_1 s) + x_2 \sin(\lambda_1 t)\sin(\mu_1 s) - x_3 \cos(\lambda_1 t)... \\ y_3 = \mathbf{w}_3^T \mathbf{x} &\propto x_1 \sin(\mu_1 s) - x_2 \cos(\mu_1 s) + x_3 * 0 + ... \end{aligned}$$

where we have the implicit requirement that the number of terms in each expansion is a multiple of 3 rather than 2 as previously. Now instead of curves we have surfaces and we can construct a three dimensional pseudo grand tour. These are the kind of curves used in the paper.

3 Finding sub-clusters

If the data set we are using has many clusters or if the dimensionality of the SOM in output space is not big enough, it may happen that after the training phase, one or several neurons were the BMUs (*Best Matching Units*) for data points belonging to different clusters: there is not a 1-1 mapping between the centres of the SOM and the clusters. Actually we do not require a 1-1 mapping but only that each centre identifies a separate class. The problem in which we are interested is the case when the SOM has not differentiated between the two clusters - the SOM has identified them as only one. However we can use the Andrews' Curves to identify the sub-clusters: the SOM has performed a crude classification perhaps finding some classes precisely but leaving others with some residual uncertainty.

To illustrate this we use a twenty dimensional artificial data set with 12 clusters (each of 85 points) and a SOM of dimensionality 3 × 3. Since, the SOM has less neurons than there are clusters in the data set, it is inevitable that some neurons form the BMU of points from more than one cluster. We create the data set placing the centres of the 12 clusters on one of 12 orthogonal axes, then we add gaussian noise in the other 8 axis and all the points were rotated using a randomly generated rotation matrix. The PCA projection of the data set can be seen in Figure 1.

As we can see in Figure 2, the SOM identifies correctly four of the clusters. But there are also two neurons that are the BMUs for two different clusters, and one neuron that is the BMU for four different clusters. We will illustrate the use of Andrews' Curves to disambiguate with this last neuron. If we use all the points for which that neuron is the BMU and we draw the Andrews' Curves (in this case we are using two slices of our surfaces) we can identify clearly the four clusters as we can see in Figure 3. If we were to use all the data points in the original data set, we would have a far less clear picture. Thus we can use Andrews' Curves to disambiguate information which contains residual uncertainty from the SOM clusters.

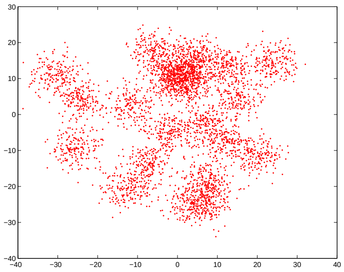

Figure 1: PCA projection of the artificial data set. Three cluster are easily identifiable, but the other 9 appear within two mixed clusters.

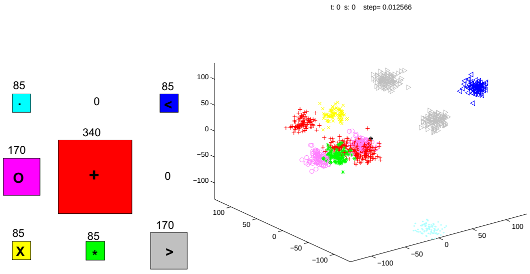

Figure 2: Left: the number of times each neuron is the BMU. Right: a projection of the data set using our surfaces.

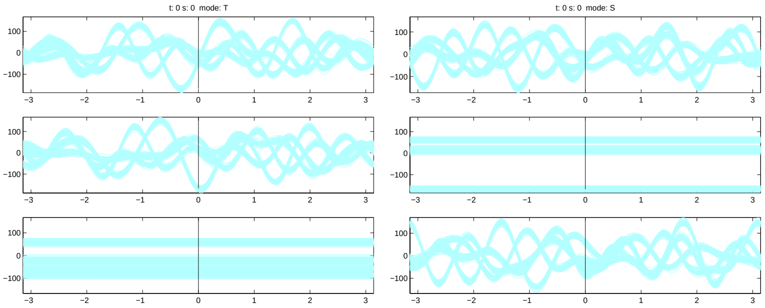

Figure 3: Left: T-slice when $s = 0$. Right: S-slice when $t = 0$. Individual sub-clusters can be identified within each plot.

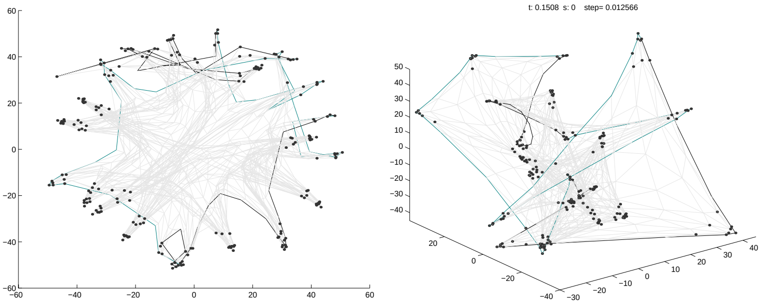

Figure 4: Left: Sammon's mapping of the prototype vector of a SOM. Right: a snapshot of the grand tour of the Andrews projections of the same SOM.

4 Clustering and projecting the SOM

There has been recent interest in using the SOM as an abstraction tool that enables us to substitute the input vectors with the model vectors as prototype vectors, and then use these to identify clusters [8]. In this way we can reduce the time needed by the clustering algorithm, since we have reduced the number of vectors to consider.

In the previous section we used the Andrews' Curves to draw the data points and try to discover sub-clusters. We can also use the Andrews' Curves to draw and project the prototype vectors. We can discover the clusters, either using the static view of the slices of our surface, or using the evolution of the projected vectors in the pseudo grand tour.

The dynamic projection is interesting on its own. It is common to use projections such as the Sammon mapping to try to visualize the structure of the SOM. But these kind of projections are static. With the use of Andrews' Curves, we can obtain a dynamic three dimensional projection and see its evolution as we change the values of s or t. The evolution of the pseudo grand tour gives us an extra dimensional perception of the projection - we can feel the fourth (time) dimension as we allow the simulation to proceed. In Figure 4 we can see the Sammon Mapping and one snapshot of the pseudo grand tour; it has to be admitted than the inevitably static projection (right diagram) as seen on the printed page does not do justice to the clarity with which the clusters jump out on the computer monitor. The SOM was trained with an artificial data set consisting of 23 clusters with centres at the vertices of a 23-dimensional hypercube plus another 52 additional dimensions with gaussian noise to get a 75-dimensional data set with 4186 observations.

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 477-482

5 Conclusion

We have, in this paper, investigated the interaction between Self Organising Maps and Andrews' Curves.

1. We have shown that Andrews' Curves can be used for visualisation of the results of data analysis by the SOM.

2. We have shown that we can perform a gross clustering of a data set using the SOM and then by selecting the data point for which a particular output neuron is the best matching unit, can iteratively find sub-clusters which the original SOM was unable to find.

3. We have discussed a new data mining technique so that, by using the interaction between the SOM, Andrews' Curves and human operators, we can walk through a data set identifying local structures in the data set.

References

[1] D. F. Andrews. Plots of high dimensional data. *Biometrics*, 28:125–136, 1972.

[2] D. Asimov. The grand tour: a tool for viewing multidimensional data. *SIAM Journal on Scientific and Statistical Computing*, 6(1):128–143, 1985.

[3] C. García-Osorio and C. Fyfe. An extension of Andrews Curves for data analysis. In *Emergent Solutions for the Information and Knowledge Economy (X SIGEF Congress)*, 2003.

[4] C. García-Osorio and C. Fyfe. Visualisation in high dimensional feature spaces. In *International Workshop on Practical Applications of Agents and Multiagents Systems (IWPAAMS 2003)*, 2003.

[5] S. Kaski. *Data Exploration Using Self-Organizing Maps.* PhD thesis, Helsinki University of Technology, 1997.

[6] T. Kohonen. *Self-Organizing Maps.* Springer, 3rd edition, 2001.

[7] J. Vesanto. SOM-based data visualization methods. *Intelligent-Data-Analysis*, 3:111–26, 1999.

[8] J. Vesanto and E. Alhoniemi. Clustering of the self-organizing map. *IEEE Transaction on Neural Networks*, 11(3):586–600, May 2000.

[9] Edward J. Wegman and Jeffrey L. Solka. On some mathematics for visualising high dimensional data. *Indian Journal of Statistics*, 64(Series A, 2):429–452, 2002.