ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 137-142

Clustering in Data Space and Feature Space

Donald MacDonald and Colin Fyfe,

Applied Computational Intelligence Research Unit,
The University of Paisley,
Scotland.

Abstract. In this paper, we extend several well known clustering methods including the neural methods known collectively as Self Organising Maps in two ways: the first uses an ϵ-insensitive version which is based on minimising the least absolute error on the data set; the second maps the data into a feature space before performing the necessary clustering. We give comparative results on astronomical data.

1 Introduction

In this paper, we compare various standard clustering methods with a recent variant of such methods which we call ϵ-insensitive clustering and a set of methods which are used in a nonlinear feature space which are known as Kernel methods.

We will not describe the basic methods for reasons of space and because they are very well known: they are k-means clustering [11], mixture of Gaussians [2], Kohonen's Self Organising Map (SOM) [8], Grossberg's ART [4] and the Generative Topographic Mapping (GTM) [3].

We first describe the new methods and then give comparative results on astronomical data.

2 ϵ-Insensitive Clustering

We have extended both the SOM and a variant known as the Scale Invariant Map (SIM) [5] with the ϵ-insensitive learning rule [6] which was developed in the context of projection networks. In this paper, we use the method with clustering networks. This changes the learning rule of these networks from being based on Least Mean Square Error to Least Absolute Error, and there is also an insensitive region in the function where learning will stop if the weight is close enough to the data point. This is useful in that, when a set of weights has almost converged, we do not have to waste computational resources on small changes which are liable to be just moving the weights slightly back and forth round the optimal values.

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 137-142

2.1 ϵ-Insensitive Self-Organizing Map

The ϵ-insensitive Self-Organising Map(ϵ-SOM) has the following algorithm.

1. Randomly select a vector from the input distribution
2. A winning weight vector, w_c, is selected which is closest of all the weight vectors to the input and the neighbourhood function, h_{ci} is calculated $\forall i$.
3. All weight vectors in the neighbourhood are updated using the learning rule

$$\Delta w_{ij} = \begin{cases} 0 & \text{if } \|x_j - w_{cj}\| < \epsilon \\ \eta.h_{ci}(\text{sign}(x_j - w_{cj})) & \text{otherwise} \end{cases} \quad (1)$$

The value of ϵ can change during learning; if it is set to zero the weight update is linear. The learning rule is less affected by outliers, since the movement of the weight towards the outlier has the same effect on the weight as the movement to a closer point which is unlike the LMS rule used in the SOM. Figure 1 shows the difference between the ϵ-SOM and the SOM on an artificial data set of three clusters; the ϵ-SOM has much fewer redundant nodes in the lattice and yet is still topology preserving in the clusters. We can see in Figure 1 that the ϵ-SOM has only 6 redundant centres (centres outside any cluster) whereas the SOM (right figure) has many redundant centres.

An interesting effect in the mapping is found by increasing the value of ϵ during learning. By setting ϵ to 0 and training the network till a reasonable mapping has been achieved then increasing the value of ϵ and continuing learning, we can move the previously redundant nodes in the map to cover an area of the data space as only the redundant nodes move. As this happens the data space will be covered by those nodes which were previously redundant, but they will still remain topology preserving, as the neighbourhood function will still adjust the position of non-redundant nodes, close to the redundant one which just moved.

An ϵ-insensitive version of the SIM is similarly defined.

3 Kernel K-means Clustering

The set of methods known under the generic title of Kernel Methods use a nonlinear mapping to map data into a high dimensional feature space in which linear operations are performed. This gives us the computational advantages of linear methods but also the representational advantages of nonlinear methods. The result is a very efficient method of performing nonlinear operations on a data set. In more detail, let $\phi(\mathbf{x})$ be the nonlinear function which maps the data into feature space, F. Then in F,we can define a matrix, K, in terms of a dot product in that space i.e. $K(i,j) = \phi(\mathbf{x}_i).\phi(\mathbf{x}_j)$. Typically we select the matrix K based on our knowledge of the properties of the matrix rather than any knowledge of the function $\phi()$. The kernel trick allows us to define

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 137-142

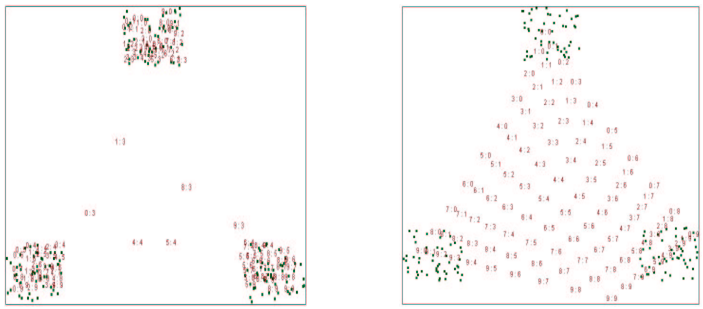

Figure 1: The left graph shows ϵ-Insensitive SOM trained on three clusters. ϵ was initially zero, and then gradually increased. There are far fewer redundant nodes and still the network is topology preserving inside and between the clusters compared with the right graph showing a SOM trained on the three cluster data set.

every operation in feature space in terms of the kernel matrix rather than the nonlinear function, $\phi()$.

In kernelising the k-means algorithm, we will follow the derivation of [12]. The aim is to find k means, $\mathbf{m}_\mu$ so that each point is close to one of the means. Now, each mean may be described as lying in the manifold spanned by the observations, $\phi(\mathbf{x}_i)$ i.e. $\mathbf{m}_\mu = \sum_i \gamma_{\mu i}\phi(\mathbf{x}_i)$. Now the k means algorithm choses the means, $\mathbf{m}_\mu$, to minimise the Euclidean distance between the points and the closest mean

$$\begin{aligned} ||\phi(\mathbf{x}) - \mathbf{m}_\mu||^2 &= ||\phi(\mathbf{x}) - \sum_i \gamma_{\mu i}\phi(\mathbf{x}_i)||^2 \\ &= k(\mathbf{x},\mathbf{x}) - 2\sum_i \gamma_{\mu i}k(\mathbf{x},\mathbf{x}_i) + \sum_{i,j}\gamma_{\mu i}\gamma_{\mu j}k(\mathbf{x}_i,\mathbf{x}_j) \end{aligned}$$

i.e. the distance calculation can be accompished in Kernel space by means of the K matrix alone.

Let $M_{i\mu}$ be the cluster assignment variable. i.e. $M_{i\mu} = 1$ if $\phi(\mathbf{x}_i)$ is in the μ^{th} cluster and is 0 otherwise. [12] initialise the means to the first training patterns and then each new training point, $\phi(\mathbf{x}_{t+1}), t+1 > k$, is assigned to the closest mean and its cluster assignment variable calculated using

$$M_{t+1,\alpha} = \begin{cases} 1 & \text{if } ||\phi(\mathbf{x}_{t+1}) - \mathbf{m}_\alpha|| < ||\phi(\mathbf{x}_{t+1}) - \mathbf{m}_\mu||, \forall \mu \neq \alpha \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

In terms of the kernel function (noting that $k(\mathbf{x},\mathbf{x})$ is common to all calculations) we have

$$M_{t+1,\alpha} = \begin{cases} 1 & \text{if } \sum_{i,j}\gamma_{\alpha i}\gamma_{\alpha j}k(\mathbf{x}_i,\mathbf{x}_j) - 2\sum_i \gamma_{\alpha i}k(\mathbf{x},\mathbf{x}_i) \\ & < \sum_{i,j}\gamma_{\mu i}\gamma_{\mu j}k(\mathbf{x}_i,\mathbf{x}_j) - 2\sum_i \gamma_{\mu i}k(\mathbf{x},\mathbf{x}_i), \forall \mu \neq \alpha \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 137-142

We must then update the mean, $\mathbf{m}_\alpha$ to take account of the $(t+1)^{th}$ data point

$$\mathbf{m}_\alpha^{t+1} = \mathbf{m}_\alpha^t + \zeta(\phi(\mathbf{x}_{t+1}) - \mathbf{m}_\alpha^t) \tag{4}$$

where we have used the term $\mathbf{m}_\alpha^{t+1}$ to designate the updated mean which takes into account the new data point and

$$\zeta = \frac{M_{t+1,\alpha}}{\sum_{i=1}^{t+1} M_{i,\alpha}} \tag{5}$$

Now (4) may be written as

$$\sum_i \gamma_{\alpha i}^{t+1}\phi(\mathbf{x}_i) \quad = \quad \sum_i \gamma_{\alpha i}^t \phi(\mathbf{x}_i) + \zeta(\phi(\mathbf{x}_{t+1}) - \sum_i \gamma_{\alpha i}^t \phi(\mathbf{x}_i))$$

which leads to an update equation of

$$\gamma_{\alpha i}^{t+1} = \begin{cases} \gamma_{\alpha i}^t(1-\zeta) & \text{for } i \neq t+1 \\ \zeta & \text{for } i = t+1 \end{cases} \tag{6}$$

This algorithm is the basis of the next two algorithms.

4 The Kernel Self Organising Map

Now the SOM algorithm is a k means algorithm with an attempt to distribute the means in an organised manner and so the first change to the above algorithm is to update the closest neuron's weights and those of its neighbours. Thus we find the winning neuron (the closest in feature space) as above but now instead of (3), we use

$$M_{t+1,\mu} = \Lambda(\alpha, \mu) \tag{7}$$

where α is the identifier of the closest neuron. Now the rest of the algorithm can be performed as before. However there is one difficulty with this: the SOM requires a great number of iterations for convergence and since $\zeta = \frac{M_{t+1,\alpha}}{\sum_{i=1}^{t+1} M_{i,\alpha}}$, this leads naturally to $\zeta \to 0$ over time. To obviate this problem, we initially select a number of centres, k and train the centres with one pass through the data set in a random order. We then have a partially ordered set of centres. We now reset all values of $M_{i,\alpha}$ to zero and perform a second pass through the data set, typically also decreasing the width of the neighbourhood function as with the normal Kohonen SOM.

To select the centre (or mean) which is closest to the grid point, we again work in feature space and calculate

$$\alpha = \arg\min_\mu ||\phi(\mathbf{x}) - \mathbf{m}_\mu||^2 \tag{8}$$

which in terms of the kernel function may be written as

$$\alpha = \arg\min_\mu \sum_i \gamma_{\mu i} k(\mathbf{x}_i, \mathbf{x}) - 2\sum_i \gamma_{\mu i} k(\mathbf{x}, \mathbf{x}_i) \tag{9}$$

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 137-142

SOM	88%
GTM	89%
Scale-invariant	72%
k-means	83%
Mixture of Gaussians	85%
ϵ-insensitive SOM	89%
ϵ-insensitive Scale-invariant	73%
Kernel ART	84%
Kernel-SOM	92%

Table 1: Classification results of the SOM, GTM, scale-invariant map, k-means, mixture of Gaussian, and ϵ-insensitive methods on extended Tholen classes of Data.

The ART algorithm may be similarly kernelised; it has the advantage that, using Gaussian kernels, a separate vigilance parameter is not necessary since the projected points all lie on the surface of a sphere since $K(\mathbf{x}, \mathbf{x}) = 1$.

5 Comparison on Astronomical Data

We now compare the clustering methods on astronomical data composed of a mixture of the 52-colour survey by Bell et al. [1] together with the 8-colour survey conducted by Zellner et al. [14] providing a set of asteroid spectra spanning 0.3-2.5mm. When this extended data set was compared by [10] to the results in Tholen [13] it was found that the additional refinement to the spectra lead to more classes in the taxonomy produced by Tholen. This was chosen as [7] showed the extended classification gave greater accuracy. The quantification of the accuracy of each of the networks was obtained by leave one out jack-knifing of the data.

We have previously compared the GTM, SOM and scale-invariant map on this data set [9]. Three networks used previously were the GTM, SOM and scale-invariant networks. We now augment these with the ϵ-insensitive clustering algorithms and the kernel versions. Results are shown in Table 1. The asteroid 308 Polyxo was consistently misclassified even when it was included in the training set. The other correct classifications show that the networks have managed to produce general classifiers.

Misclassification is counted where two or more asteroid types are assigned to one node (with the GTM a misclassification is where the means of two classes are within a predefined radius - 0.05) . We see that the ϵ-insensitive methods show a slight improvement over the standard methods but the kernel methods make a huge improvement.

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 137-142

References

[1] J. F. Bell, P. D. Owensby, B.R. Hawke, and M.J. Gaffey. The 52 colour asteroid survey: Final results and interpretation. *Lunalr Planet Sci. Conf, XiX*, pages 57–58, 1988.

[2] C. Bishop. *Neural Networks for Pattern Recognition.* Oxford:Clarendon Press, 1995.

[3] C. Bishop, M. Svensen, and C.K.I Williams. Gtm: a principled alternative to the self-organizing map. In *Proceedings International Conference on Artificial Neural Networks*, pages 165–170. Springer-Verlag, 1996.

[4] G. Carpenter and S. Grossberg. *Pattern Recognition by Self-Organizing Neural Networks.* The MIT Press, 1991.

[5] C. Fyfe. A scale invariant feature map. *Network: Computation in Neural Systems*, 7:269–275, 1996.

[6] C. Fyfe and D. MacDonald. Epsilon-insensitive hebbian learning. *Neurocomputing*, 2001. (Accepted for publication).

[7] E. S. Howell, E. Merenyi, and A. Lebofsky. Classification of asteroid spectra using a neural network. *Journal of Geophysical Research*, 99(E5):10847–10865, May 1994.

[8] T. Kohonen. *Self-Organisation and Associative Memory.* Springer-Verlag, New York, 1988.

[9] D. MacDonald, S. McGlinchey, J.Kawala, and C. Fyfe. Comparison of kohonen, scale-invariant and gtm self-organising maps for interpretation of spectral data. In *Seventh International Symposium on Artificial Neural Networks, ESANN99*, pages 117–122, 1999.

[10] E. Merenyi. Self -organising anns for planetary surface composition research. *Journal of Geophysical Research*, 99(E5):10847–10865, 1994.

[11] D. Michie, D. J. Spiegelhalter, and C. C. Taylor, editors. *Machine Learning: Neural and Statistical Classification.* Ellis Horwood, 1994.

[12] B. Scholkopf, A. Smola, and K.R. Muller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

[13] D.J. Tholen. *Asteroid Taxonomy from Cluster Analysis of Photomerty.* PhD thesis, University of Arizona, Tucson, 1984.

[14] B. Zellner, D.J. Tholen, and E.F. Tedesco. The eight colour asteroid survey: Results from 589 minor planets. *Icarus*, pages 355–416, 1985.