
Data Mining II, C.A. Brebbia & N.F.F. Ebecken (Editors)
© 2000 WIT Press, www.witpress.com, ISBN 1-85312-821-X

Hierarchical clustering for data mining by rbf network

Ö. Ciftcioglu & S. Sariyildiz
Delft University of Technology
Faculty of Architecture, Department of Informatics
Delft, The Netherlands

Abstract

Clustering is one of the dominant techniques of exploratory data analysis. In the context of data mining, one of the powerful clustering methods is the method of orthogonal least squares (OLS) applied to radial basis functions (RBF) network. However, for data mining case, conventional utilization of OLS learning is not desirable even though the learning process feasible for an amount of data at hand. This is due to the fact that some dominant associations in the data can easily obscure many interrelations of interest among the data at hand and also because of this, the rest of the associations for identification can heavily be limited. Therefore, as novel RBF clustering for data mining, the data set is separated into several subsets so that each subset is subjected to clustering separately. The dominant clusters in each subset are accumulated throughout the modular processing of total data set. The newly formed reduced data set which comprises the patterns from the clusters of the subsets, is subjected to final RBF network clustering by OLS for hierarchical cluster gradation of the subsets to identify the existing interrelations being searched for. In this way the associations are discovered in a perspective without obscurity caused by some strong but local associations in the given data set. The paper deals with the details of the novel RBF clustering for data mining.

1. Introduction

Clustering is one of the dominant techniques of exploratory data analysis. There are a number of ways for clustering [1]. They can broadly be divided into two categories as clustering with unsupervised and with supervised learning. In the unsupervised case, the clustering is performed where given a data set and the

© 2000 WIT Press, www.witpress.com, ISBN 1-85312-821-X

distance function, one starts with the number of clusters assigning each pattern to a separate cluster and proceeds to merge the clusters that are the closest. In the supervised case, we are concerned with a collection of labeled data that come in the form of ordered pairs. This can be seen as a feature vector describing the data and its class assignment. Eventually, the associations between the feature vectors and the class assignments are established. In the context of data mining perhaps the major disadvantage of unsupervised clustering is due to that the clusters are established from the input data and the final associations with some output functions are performed afterwards. In this case the associations may not be optimal since the established clusters may not be the optimal representations with respect to associations being looked for. In contrast with this, in the supervised case the clustering and associations are performed in a constructive way so that the most effective representation of associations is obtained. In the context of data mining, one of the powerful supervised clustering method is the method of orthogonal least squares (OLS) [2] applied to radial basis functions (RBF) network where the clusters are hierarchically ordered according to their effectiveness of representations for the associations. The Introduction of fuzzy computing or fuzzy associative memory (FAM) [3] matrices for data mining can easily extend these associations to linguistic variables by fuzzy logic techniques. As a variant of feed-forward neural networks, RBF networks are widely treated in the literature in the last decade so that their potential for data mining is well recognized today. In this research, a novel RBF clustering for data mining is dealt with. The organisation of the paper is as follows. Section 2 gives the brief description of RBF networks and associated OLS algorithm for training. Section 3 describes the hierarchical RBF network training. Section 4 gives the results of data mining with RBF clustering which is followed by conclusions.

2. RBF Network and the OLS algorithm

A basic radial basis function (RBF) network may be depicted as shown in Fig.1. Without loss of generality, the number of outputs in the network can be extended to a multi-output case. The architecture consists of an input layer, a hidden layer and an output layer. The hidden layer consists of a set of radial basis functions as nodes. Each node has a parameter vector $\mathbf{c}$ defining a cluster centre dimension of which is equal to the input vector. The hidden layer node calculates the Euclidean distance between the centre and the network's input vector. The distance calculated is used to determine the radial base function output. Conventionally, all the radial basis functions in the hidden layer nodes are the same type and usually gaussian. The response of the output layer node(s) can be seen as a map f: $R^n \rightarrow R$, of the form

$$f(\mathbf{x}) = \Sigma\, \mathbf{w}_i\, \Phi(||\mathbf{x} - \mathbf{c}_i||)^2 \qquad (1)$$

Here the summation is over the number of training data N. $\mathbf{c}_i$ (i=1,2,….,N) is the i-th centre which may be equal to the input vector $\mathbf{x}_i$ or may be determined in some

Data Mining II, C.A. Brebbia & N.F.F. Ebecken (Editors)
© 2000 WIT Press, www.witpress.com, ISBN 1-85312-821-X

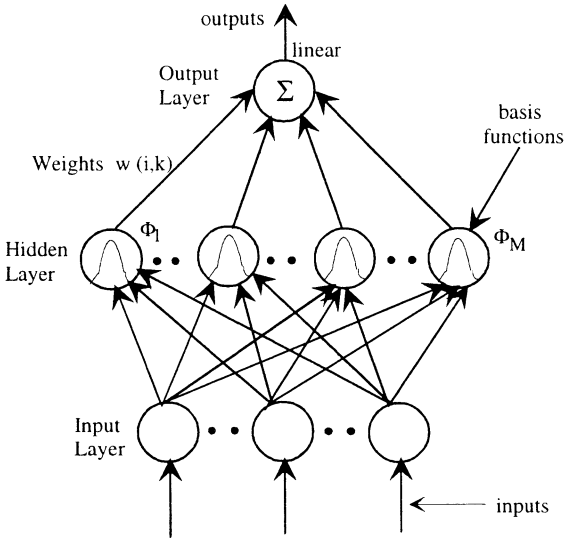


Figure 1: *A basic RBF network architecture*

other way. Once the basis function outputs are determined, the connection weights from hidden layer to the output are determined from a linear set of equations. As a result, accurate functional approximation is obtained. The complexity increases as the size of the training data increases. For a large data set this may become unpractical. Therefore it is desirable to use limited number of hidden layer nodes in place of having a number equal to N. In the present description for sake of simplicity in representation and description, a single function is considered so that neural network has one output for each multivariable input. For this case the output is given by

$$f(x) = w_o + \sum_{j=1}^{N} w_i \phi(\| x - c_j \|) \tag{2}$$

where $x \in R^p$ is the input vector . For the training the network, orthogonal least squares method is used due to its outstanding features for this study. Essentially RBF network is the generalization of the linear regression model

$$d(t) = \sum_{i=1}^{N} w_i \, r_i(t) + \varepsilon(t) \tag{3}$$

corresponding to the preceding equation. The desired output, $r_i(t)$ are the regressors which are some fixed functions of x(t). If the model is right, then ε is not correlated with the regressors. For a set of input out pairs, this model, in matrix form can be expressed as

$$\mathbf{d} = R\,\mathbf{w} + E \tag{4}$$

where **d**, is the desired output vector; R is the regression matrix that consists of regressor vectors r_i each of which has dimension of M which can extend to N, **w** is

© 2000 WIT Press, www.witpress.com, ISBN 1-85312-821-X

the parameter vector; E is the error vector. The regressor vectors r_i form a set of basis vectors, and the least squares estimation of **w** provides that (R**w**) is the projection of **d** onto the space spanned by these vectors. The OLS method makes orthogonal decomposition of the R matrix of the form

$$R=W\,A,\quad A=\begin{bmatrix} 1\ a_{12}\ \ldots. & a_{1M} \\ 0\ 1\ \ldots\ldots & a_{2M} \\ 0\ 0\ 0\ldots\ldots & \\ \ldots. & a_{M-1,M} \\ 0\ 0\ \ldots\ldots\ldots\ldots 0\ \ 1 & \end{bmatrix}_{M\times M} \tag{5}$$

and W is an N×M matrix with orthogonal columns such that

$$W^TW=H,\qquad H=\begin{bmatrix} h_1\ \ldots\ldots 0 \\ \ldots\ldots\ldots\ldots \\ 0\ldots\ldots\ h_M \end{bmatrix}_{M\times M} \tag{6}$$

H is a diagonal matrix. dimension of MxM. With the definitions

$$\mathrm{W=R\ A^{-1}}\ ,\qquad \mathbf{q}=\mathrm{A}\ \mathbf{w} \tag{7}$$

the matrix equation **d** = R **w**+ E , in this case takes the form

$$\mathbf{d}=\mathrm{W}\ \mathbf{q}+\mathrm{E}, \tag{8}$$

which has the solution for the estimation of the transformed parameter vector q as

$$\hat{q}=H^{-1}W^T\,d\quad or\quad \hat{q}=\frac{w_i^T d}{w_i^T w_i}\ ,\quad i=1,\ldots.M \tag{9}$$

There is an efficient orthogonalization procedure for getting W. The traditional Gram-Schmidt computes one column of A at a time and orthogonalizes R at the same time. Namely, we add matrix W to both sides of the matrix equation **R=WA**, we obtain **W=R-W (A-I)**. Explicitly

$$[w_1\ldots\ldots w_M]_{NxM}=[r_1\ldots\ldots r_M]_{NxM}-$$

$$[w_1\ldots\ldots w_M]_{NxM}\begin{bmatrix} 1\ a_{12}\ \ldots. & a_{1M} \\ 0\ 1\ \ldots\ldots & a_{2M} \\ 0\ 0\ 0\ldots\ldots & \\ \ldots. & a_{M-1,M} \\ 0\ 0\ \ldots\ldots\ldots\ldots 0\ \ 1 & \end{bmatrix}_{M\times M} \tag{10}$$

Now, the vectors $\mathbf{w}_i$ and a_{ij} can be computed iteratively by the method of Gram-Schmidt. In this method, at the k-th step, first the k-th column of matrix A and afterwards the vector $\mathbf{w}_k$ are determined, $\mathbf{w}_k$ being orthogonal to the preceding k-1 columns where k = 1, ..., M. Finally $\mathbf{d}=W\ \mathbf{q}$ is computed. As soon as $\mathbf{q}$ is determined, the actual weights $\mathbf{w}$ are computed by

$$\mathbf{w} = A^{-1}.\ \mathbf{q} \tag{11}$$

In terms of energy transmitted from input to the output, we can write the following algebraic balance equation for zero-mean output vector:

$$d^T d = \sum_{i=1}^{M} q_i^2 w_i^T w_i + E^T\ E \tag{12}$$

Normalizing by $\mathbf{d}^T\mathbf{d}$ we obtain the relative energy contribution from each hidden node, i.e. from each regressor to the output as

$$z_i = \frac{q_i^2 w_i^T w_i}{d^T d} \qquad i = 1,....M \tag{13}$$

3. Hierarchical clustering for data mining

Conventionally, the main drawback of training RBF networks with OLS method with large volume of data is that the learning process becomes formidably cumbersome since each pattern forms a cluster center subject to a hierarchical gradation. Therefore, the dimension of the square matrix A in eq.11 becomes formidably large for computing. However, next to the training inconvenience of large volume of data, for data mining OLS learning using large volume of data is not desirable. This is due to the fact that some individual dominant associations in the data can easily obscure many interrelations of interest among the input and output data at hand. As result of this the rest of the associations for identification can heavily be limited. To circumvent this, the large data set can be divided into several subsets and each subset can be subjected to OLS gradation and limited number of patterns from the uppermost importance can be selected. After this procedure having been done for each subset, the total number of newly composed patterns is subjected altogether again to a final OLS training procedure. In this way, the number of nodes finally selected are the representatives of the total data, i.e., the total number of patterns, while total number of nodes can be so arranged that it is much less then the total number of patterns, i.e., $M << N$. For instance, let us assume N=1000. Each subset can be selected as N_i=100 and M_i=10, (i=1,..,10) so that the eventual hierarchically designed RBF network is formed for N=1000 and $1 \le M \le 100$.

Referring to preceding hierarchical process described, in this research, as novel RBF clustering for data mining, the data set is separated into several subsets so that each subset is subjected to OLS clustering separately. The dominant clusters in each subset are accumulated throughout the total data set. The newly formed

reduced data set which comprises the patterns from the clusters of the subsets, is subjected to final RBF clustering by OLS for hierarchical cluster gradation to identify the existing interrelations being searched for. In this way the associations are discovered in a perspective without probable obscurity caused by some strong but local associations in the given data set of large volume of data. Conspicuously, the data mining training virtually requires no limitation for the number of patterns in use, for all practical purposes.

One of the applications for hierarchical data mining is the efficient and effecrive wavelet transform by RBF network. The accurate wavelet transform by RBF networks has already be reported earlier[4]. Very briefly, discrete wavelet transform is similar to dicrete Fourier transform in the sense that a bock of time series data is tranformed into another block of data of the same size. In the case Fourier transform we obtain frequency spectrum of the data. In the case of wavelet transform [5,6] we obtain coefficients that are called wavelet coefficients and they are used for time-frequency representation of the time series data. Although, RBF approach can provide accurate wavelet transform, it uses relatively high number of basis functions. In real time applications, the real added value of soft computing approach for wavelet transform is due to the speed of computation. However, for high number of basis functions the speed of computation with RBF network might be comparable with actual computation using a fast transform algorithm. Therefore, parsimony in the use of number of basis functions is important. For wavelet transform by soft computing, to train the RBF network originally the number of patterns used is 100 and the number of basis functions employed was relatively high to achieve high accuracy. In this research, for the same transform task the training of RBF network is carried out using larger number of data. In this way more representative centers of radial basis functions are identified from a larger set of patterns relative to a smaller data set used earlier. By doing so, the anticipated result is that the number of centres subject to transform computation for the same accuracy in multivariable functional approximation can be reduced.

4 Experimental research

To illustrate the effect of the hierarchical clustering total 500 blocks of time series pump data is considered. Each block contained 32 samples. Each block is referred to as pattern here. In the first case, only first hundred is used for RBF network training. The network contained 32 inputs for the time series data and 32 outputs as the corresponding transforms. The number of basis functions was selected to be as low as 20. Following the OLS training, as a test the trained network is used to estimate the wavelet transform of 100 consecutive blocks of time series data. Each block has a length of 32 and belongs to the last hundred blocks of the total number of blocks. This means, the test patterns are never seen by the network before. The results are shown in Figure 2. The same experiment was repeated with the same experimental conditions ,i.e., 20 basis functions with OLS training using hierarchical clustering and ensuing RBF network's weight identification, where 400 patterns with a length of 32 were used for this case. The results obtained with

© 2000 WIT Press, www.witpress.com, ISBN 1-85312-821-X

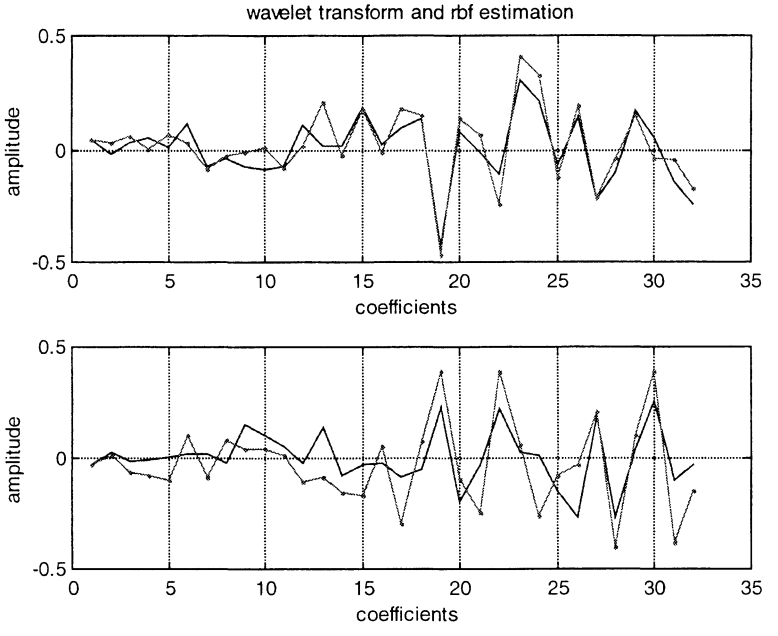


Figure 2: RBF network trained by OLS for wavelet transform estimation. The number of patterns used is 100. Estimations are indicated by dots

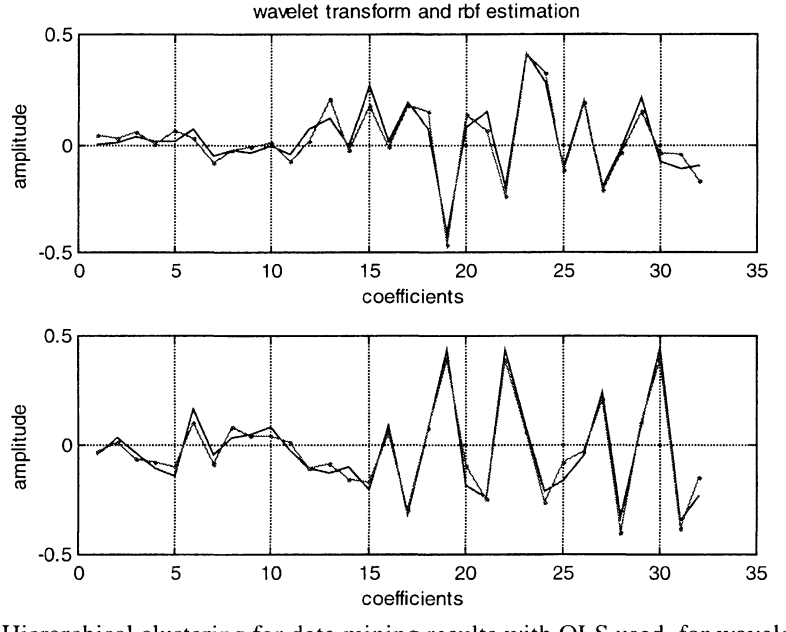


Fig.3. Hierarchical clustering for data mining results with OLS used for wavelet transform estimations. The number of patterns used is 400 Estimations are indicated by dots

© 2000 WIT Press, www.witpress.com, ISBN 1-85312-821-X

the same test data, that is using the last hundred of the total number of blocks, are shown in Figure 3. From the comparison of Figure 2 with Figure 3, the improved estimation by hierarchical clustering is observed as this was expected. Basically, in the latter case, more representative basis functions are selected out of 400 compared to 100 in the former case represented in Figure 2.

Similar analysis is carried out using 100 patterns containing also the patterns shown in Figs.2 and 3. In this case the training results are given in Figure 4 where the patterns subjected to estimation here were seen by the network before. This is in contrast with the cases represented by figures 2 and 3.

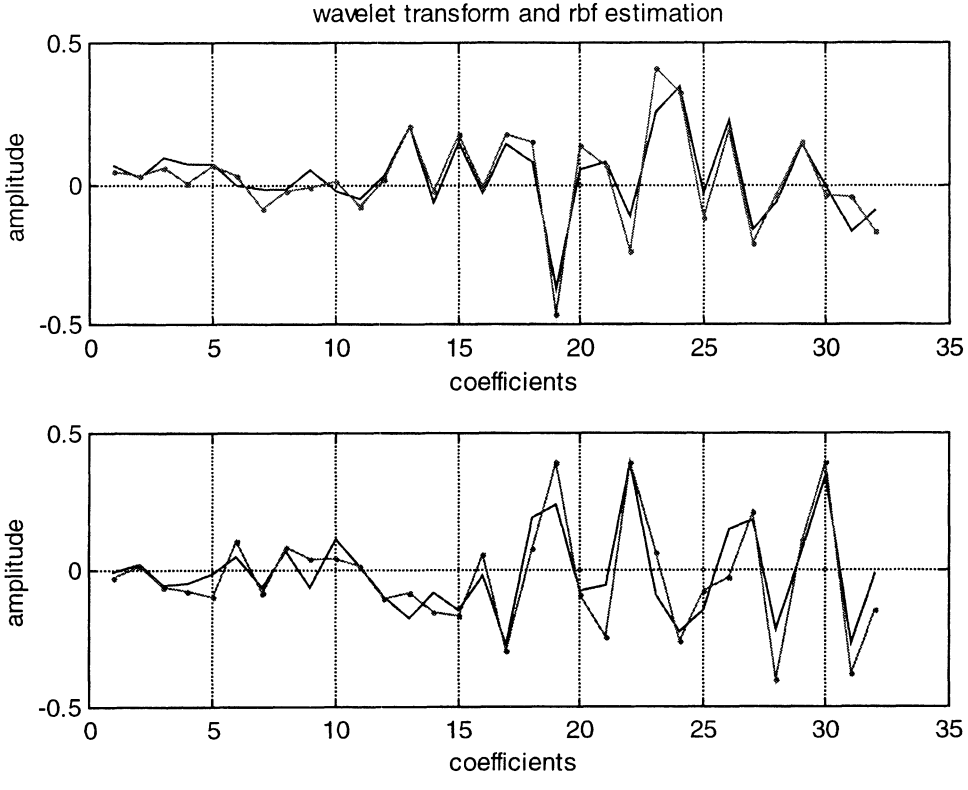


Figure 4: RBF network trained by orthogonal least squares method for wavelet transform estimation. The number of patterns used is 100. Estimations are indicated by dots

Comparison of fig.2 and fig.4 reveals that, even the estimation by the network trained by hierarchical clustering is slightly better than that obtained by the network where the training data included the patterns subjected to comparisons through figures 2 to 4. The result is rather conspicuous indicating the effectiveness of the hierarchical clustering for data mining.

5 Discussion and conclusions

The method of OLS considers input as well as output spaces for clustering. This is an important advantage against unsupervised clustering algorithms where only input space is considered. This is simply due to taking into account the associations between input space and output space during clustering where the associations can require certain modification of the cluster centres depending on the nature of the associations. To achieve the same thing in unsupervised clustering case, however, output space can be concatenated to input space and one of the unsupervised clustering methods can be performed. In this case, the comparison of the results obtained from unsupervised and supervised clustering would be interesting and is subject to another study on this line.

The presented hierarchical OLS clustering can be seen as modular RBF network approach for OLS clustering so that the limitations on the utilisation of RBF networks due to size of the data is greatly alleviated. In the terminology of neural networks, a network is said to be modular if the computation performed by the network can be decomposed into two or more modules that operate on distinct inputs without communicating with each other. Next to modularity, basically, the gradation of the clusters is an important asset of OLS clustering that it can be used for the gradation of associations between two sets of data where each set belongs to a separate space designed as input space and output space. One application area in this respect is the architectural design decision support. To conceive this, we consider a number of architectural design aspects of a certain architectural design to be achieved. Each design aspect requires a number of related considerations and therefore each aspect as a pattern in the input space can be related to a certain design outcome with its associated design features as another pattern at the output space. For instance one design aspect at the input space might be safety, and the associated feature might be cost at the output space. The linguistic architectural quantities might be dealt with fuzzy logic [7] techniques for subtle pattern formations. Such an architectural data mining application has been reported before [8].

This paper has studied the modular use of the orthogonal least squares radial basis functions network training as clustering algorithm for data mining purposes. Practically, the proposed hierarchical clustering at the same time represent a robust information processing system for classification and data fusion. It was pointed out that, it provides performance advantages and therefore useful to a broad spectrum of application domains that might span engineering disciplines, as exemplified in the paper, and Architecture.

© 2000 WIT Press, www.witpress.com, ISBN 1-85312-821-X

6 References

[1] Bezdek, J.C., *Pattern recognition with fuzzy objective function algorithms*, Plenum, New York, 1981

[2] Chen S, C.F.N. Cowan and P.M. Grant, *Orthogonal Least Squares Algorithm for Radial Basis Function Networks*, IEEE Trans. on Neural Networks, Vol.2, No.2, March, 1991

[3] Kosko, B. *Neural Networks and Fuzzy Systems*, Prentice Hall, Englewood Cliffs, New Jersey , 1992

[4] Ciftcioglu Ö., *Wavelet transform by soft computing* in 'Advances in soft computing' by R. John and R. Birkenhead (Eds.), Physica-Verlag, A Springer-Verlag Company, Heidelberg, New York, 2000

[5] Mallat S. *Wavelet tour of signal processing*, Academic Press, San Diego, London, 1999

[6] Daubechies I, *Ten lectures on wavelets*, CBMS-NSF regional conference series in applied mathematics Vol.61, SIAM, 1992

[7] Zadeh, L.A., *Outline of a new approach for the analysis of complex systems and decision processes*, IEEE Trans. on Systems, Man and Cybernetics, Vol. SMC-3, No.1, January 1973

[8] Durmisevic S. Ö. Ciftcioglu and S. Sariyildiz, *An application of neural* network in post-occupancy evaluation of underground stations, Proc. Construction Information Technology CIT2000 , June 28-June 30, Reykjavik, Iceland, *2000*