RULE-EXTRACTION FROM RADIAL BASIS FUNCTION NETWORKS

Kenneth J. McGarry, John Tait and Stefan Wermter

School of Computing, Engineering and Technology,
St Peters Campus, St Peters Way,
University of Sunderland, Sunderland, England, SR6 ODD
email: cs0kmc@cis.sunderland.ac.uk
Content Area: Cognitive Modeling

ABSTRACT

Radial basis neural (RBF) networks provide an excellent solution to many pattern recognition and classification problems. However, RBF networks are also a localist representation technique that enables the easy conversion of the hidden units into symbolic rules. This paper examines rules extracted from RBF networks. We use the iris flower classification task and a vibration diagnosis clasification task to illustrate the new knowledge extraction techniques. The rules are analyzed in order to gain knowledge and insight into the network representations. We argue that the localist Gaussian representation in RBF networks is particulary useful for rule extraction.

INTRODUCTION

The work described in this paper is concerned with rule extraction from radial basis function (RBF) networks. Rule extraction is recognised as a powerful technique for neurosymbolic integration within hybrid systems [22, 10]. Radial basis function networks are a localist type of learning technique [2, 12] and have been applied to several real-world, large-scale problems of considerable complexity [23]. They are good at pattern recognition and are robust classifiers, with the ability to make decisions about imprecise input data [18]. Furthermore, they offer robust solutions to a variety of classification problems such as speech, character and signal recognition, as well as functional prediction and system modeling where the physical processes are not understood or are highly complex [8].

Because of these advantageous properties and the possible support of rule extraction by localist representations we focus on RBF networks. Localist learning systems generally contain elements that are responsive to only a limited section of the input space. This is quite different from the distributed approach of multi-layer perceptron networks (MLP) [9]. The local nature of RBF networks makes them an interesting platform for performing rule extraction. Here we examine the ability of a new rule extraction algorithm to extract meaningful rules that describe the overall performance of a particular RBF network. The research, carried out on the extracted rule quality and complexity, also has a direct bearing on the use of rule extraction algorithms for data mining and knowledge discovery.

The data sets we used comprised a benchmarking data set namely, Fisher's iris set and a real-world condition monitoring data set. The iris data set consists of three classes of flower with 50 patterns each. One class is linearly separable while the other two are not. The condition monitoring data set is a fault diagnosis problem. The data set consists of 10 input features and seven output classes with several hundred patterns in each class representing the recognized errors.

This paper is structured as follows: Section two outlines in a general way the benefits and techniques currently available for rule extraction. Our rule extraction technique for RBF networks are presented in more detail. Section three describes the architecture and features of radial basis function networks. Section four discusses the results and implications of the experiments. Section five is concerned with related work on rule extraction.

RULE EXTRACTION FROM NEURAL NETWORKS

In this section we discuss motivations, techniques and methodology for rule extraction from RBF networks. RBF networks provide a localized solution [2, 12] that is amenable to rule extraction. Previous work on extracting rules from radial basis functions [20] has investigated generating probabilistic rules or has identified certain neuro-fuzzy similarities [7]. This research will be discussed in de-

tail in section five.

Rule extraction has been carried out upon a variety of neural network types such as multi-layer perceptrons [19, 3, 5], Kohonen networks [21] and recurrent networks [14]. The advantages of extracting rules from RBF neural networks are:

- The knowledge learned by a neural network is generally difficult to understand by humans. The provision of a mechanism that can interpret the networks input/output mappings in the form of rules would be very useful.

- Deficiencies in the original training set may be identified, thus the generalization of the network may be improved by the addition/enhancement of new classes. The identification of noisy training data for removal would also enhance network performance.

- Analysis of previously unknown relationships in the data. This feature has a huge potential for knowledge discovery/data mining and possibilities may exist for scientific induction.

- Once having extracted rules from a neural network we have a rule base that has the potential to be inserted back into a new network with a similar problem domain.

RADIAL BASIS FUNCTION NETWORKS

Radial basis function (RBF) neural networks were independently proposed by a number of researchers [4], [11] and they have been proved to be capable of universal function approximation [15]. Figure 1 shows the architecture of a typical RBF network.

The RBF network consists of feedforward architecture with an input layer, a hidden layer of RBF units and an output layer of linear units. The response of the output units is calculated using equation 1.

$$\sum_{j=l}^{J} W_{lj} Z_j(x) \qquad (1)$$

where:
W = weight matrix
Z = hidden unit activations
x = input vector

The input layer simply transfers the input vector to the hidden units, which form a localized response to the input pattern. The activation levels of the output units provide an indication of the nearness of the input vector to the classes. Learning is normally undertaken as a two-stage process. An unsupervised clustering technique is appropriate for the hidden layer while a supervised method is applied to the output layer units. The nodes in the hidden layer are implemented by kernel functions, which operate over a localized area of input space. The effective range of the kernels is determined by the values allocated to the centre and width of the radial basis function. The Gaussian function is very appropriate for rule extraction and has a response characteristic determined by equation 2.

$$Z_j(x) = exp\left(-\frac{||x-\mu||^2}{\sigma_j^2}\right) \qquad (2)$$

where:
μ = n-dimensional parameter vector
σ = width of receptive field
x = input vector

The output of a hidden unit is radially symmetric in input space. Therefore a hidden unit will give an output dependent upon the Euclidean distance between the centre of the basis function and the input vector. RBF networks are an appropriate choice for both classification tasks and function approximation. The adjustable parameters within a radial basis function network that effect classification accuracy and that may provide information for rule extraction are:

- Number of basis functions used.

- Location of the centre of the basis function.

- Width of the basis function.

- Weights connecting the hidden RBF units to the linear output units.

Rule extraction algorithm

Our algorithm developed to extract rules used a straightforward approach. The input weight space was summarised in terms of maximum and mimimum values per input dimension. The extracted rule set is compact, providing one rule for each output class. Figure 2 describes the extraction algorithm in detail.

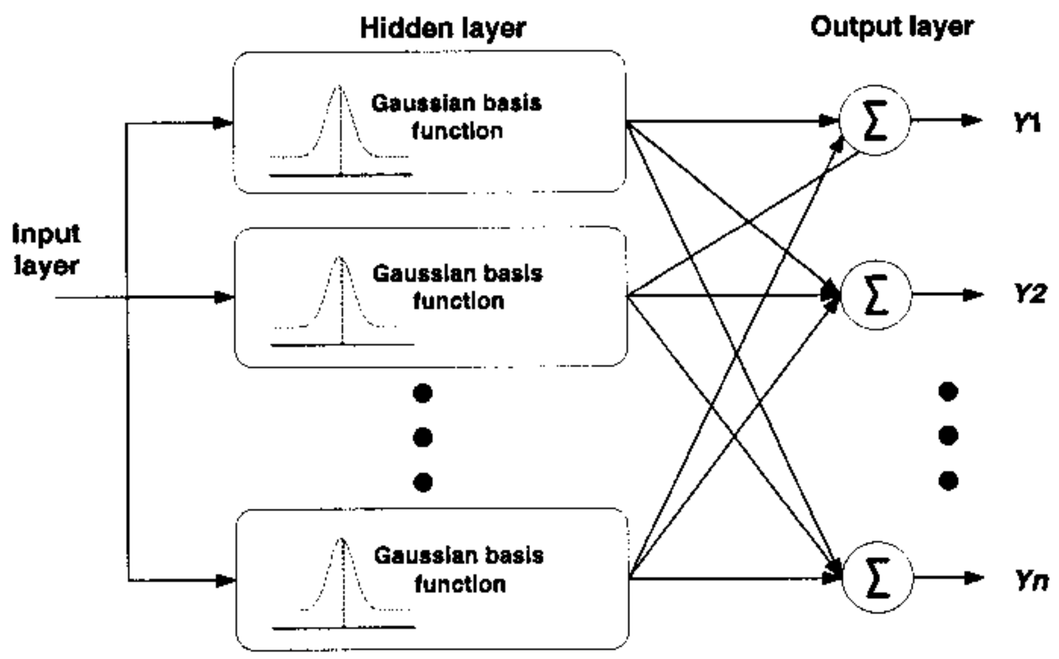

Figure 1: Radial basis function network

Input:
Hidden weights μ (centre positions)
Output:
One rule per output class
Procedure:
Train RBF network on data set
Cluster hidden units by class
For each class cluster
For each μ
Get min value
Get max value
For each class
Write out rule by:
For each μ = [min - max] interval
Join intervals with AND
Add Class label
Write rule to file

Figure 2: Rule extraction algorithm

The algorithm analyzes the RBF network and determines which hidden units contribute to the identification of the various classes. This is achieved by observing the hidden unit activations during network testing. The RBF network is presented with patterns where the class identity is known. Those hidden units that have activations between 0.5 and 1.0 are said to be contributing significantly to the identification of that particular class. Each cluster of hidden units assigned to a class have their cluster centre weights μ analyzed for maximum and minimum ranges. This procedure ensures that a range of minimum and maximum values is assigned to each dimension of the input space. This range guarantees that only valid rules may be formed. Each rule consists of the minimum and maximum values acting as valid ranges for each antecedent and only one rule per output class is generated.

EXPERIMENTAL RESULTS

The experimental work consisted of rule extraction from two main data sets. The iris data and an industrial fault diagnosis task.

Iris data set

The first data set used was Fishers iris data [6]. The iris data set is well known within the neural network community as a benchmark to demonstrate the effectiveness of new algorithms. The iris data set consists of three classes of flowers with 50 patterns each, one class is linearly separable (Setosa) while the other two are not (Versicolor and Virginica). The data is continous valued with a dimensionality of four. The inputs correspond to the plant features such as the sepal length, sepal width, petal length and petal width. See table 1. A network was trained with 50 RBF units, each unit represents a single epoch through the training data.

Vibration data test set

The second data set consisted of spectral vibration data gathered from a fault diagnosis application. Many large items of machinery are regulary monitored by analyzing the spectral vibration data that is generated when they are operating. The vibration signatures produced are very distinct and any changes in these patterns may

Table 1: Examples of Iris data set. Where: SL=sepal length,SW=sepal width,PL=petal length,
PW=petal width,C1=Setosa,C2=Versicolor and C3=Virginica.

Input features				**Output classes**		
SL	SW	PL	PW	C1	C2	C3
5.2	3.5	1.5	0.2	1	0	0
6.2	2.8	4.8	1.8	0	0	1
6.7	3.1	4.4	1.4	0	1	0
5	3	1.6	0.2	1	0	0
6.7	3.3	5.7	2.1	0	0	1

be used to detect faults in the mechanical condition.

The data set consists of 681 samples composed of 10 input features and seven output classes. The input features are continuous values representing the spectral interpretation of the running speed (rotations per minute (RPM)) of the motor or fan and the various harmonics that occur at twice, three times running speed etc. This information is normally derived from a fast fourier transform (FFT) and is often measured in acceleration $(mm/sec)^2$. The data represents several fault conditions but also includes healthy data. Several aspects of the data are highly non-linear and linearly non-separable. Table 2 presents a sub-set of the vibration data test set. A network consisting of 250 hidden RBF units was trained.

Table 2: Examples of vibration spectra data set. Where: RPM1=motor speed, RPM2=twice motor speed

Input features			**Faults**		
RPM1	RPM2	RPM3	F1	F1	F2
3.205	1.687	1.046	1	0	0
1.399	1.273	1.328	0	1	0
0.843	1.017	0.755	0	1	0
0.406	0.457	0.289	0	0	1

ANALYSIS OF RESULTS

The extracted rules provided reasonable classification results on the Setosa class of the iris dataset. The rules were tested against the previously unseen test data. The extracted rules for the iris data set and vibration diagnosis problem are shown in figure 3 and figure 4 respectively.

Rule 1
IF (SL $\geq$ 4.4 AND $\leq$ 5.7) AND
IF (SW $\geq$ 2.9 AND $\leq$ 4.4) AND
IF (PL $\geq$ 1.3 AND $\leq$ 1.5) AND
IF (PW $\geq$ 0.2 AND $\leq$ 0.4)
THEN..Setosa

Rule 2
IF (SL $\geq$ 4.9 AND $\leq$ 6.9) AND
IF (SW $\geq$ 2 AND $\leq$ 3.1) AND
IF (PL $\geq$ 3.5 AND $\leq$ 5) AND
IF (PW $\geq$ 1 AND $\leq$ 1.7)
THEN..Versicolor

Rule 3
IF (SL $\geq$ 5.8 AND $\leq$ 7.2) AND
IF (SW $\geq$ 2.8 AND $\leq$ 3.1) AND
IF (PL $\geq$ 4.5 AND $\leq$ 5.8) AND
IF (PW $\geq$ 1.5 AND $\leq$ 2.4)
THEN..Virginica

Figure 3: Extracted rules from iris data set

Rule 1 describes the valid intervals an input vector must conform to, so that a Setosa class is recognised. Further examination of the other rules shows that multiple classes may be activated. It is interesting to note in the case of the vibration rule number 4, that certain antecedents have very low values. From experience we know that such parameters have no value in classifying the specific fault. Therefore, to a limited degree the algorithm can prune unnecessary antecedents.

Rule 4
IF (RPM $\geq$ 0.41821 AND $\leq$ 3.0124) AND
IF (2RPM $\geq$ 0.45858 AND $\leq$ 2.8013) AND
IF (3RPM $\geq$ 0.52502 AND $\leq$ 1.6375) AND
IF (4RPM $\geq$ 0.44754 AND $\leq$2.2021) AND
IF (5RPM $\geq$ 0.23192 AND $\leq$ 0.86202) AND
IF (HarmPow $\geq$ 2.3951 AND $\leq$ 14.6983)
AND
IF (IRD $\geq$ 62.1835 AND $\leq$ 248.734) AND
IF (ORD $\geq$ 0.0001 AND $\leq$ 0.0001) AND
IF (Train $\geq$ 0.0001 AND $\leq$ 0.0001) AND
IF (Ball $\geq$ 0.0001 AND $\leq$ 0.0001)
THEN..IRD

Figure 4: Extracted rules from vibration data set

However, the rules extracted from the vibration analysis domain proved to be less effective at describing the networks perfor-

mance. However, for humans to understand the extracted rules it is essential that only a small number of key rules are generated. The rules represent local solutions that must be organized into clusters representing the global trends and relationships within the data. Such a rule extraction procedure is currently under development and is similar in scope to the techniques proposed by Andrews and Geva [1].

The accuracy of the rules for the iris data was 40% with 3 rules generated. The vibration data produced 7 rules with an overall accuracy of 25%. The number of rules generated is simply the number of classes in the dataset. The original RBF network trained on the iris data had an accuracy of 88% and had an accuracy of 75% when tested on the vibration data. The network in both cases was tested using 50% of the data set. In general, there is a trade-off between a large number of extracted rules and a high accuracy versus a small number of rules and a low accuracy.

DISCUSION AND CONCLUSION

Previous work involving rule extraction and inserting prior knowledge into locally responsive units used a Bayesian framework which enabled the incorporation of an inductive bias [17]. Other approaches have considered the problem in terms of inserting symbolic rules into a RBF network [1, 20].

The feasibility of knowledge transfer has also been investigated with multi-layer perceptron networks [13]. The main problem with multi-layer perceptron networks is to identify those hyperplane positions that may be of use for the second task. This process is difficult due to the distributed nature of multi-layer perceptron networks.

The solution proposed by Pratt [16] involves the use of information theory applied to the decision boundaries formed by the hidden unit hyperplanes. Those boundaries with a high mutual information measure are useful in class separation and are transferred to the target network.

Localist neural network representations such as radial basis functions are well suited for symbolic rule extraction and knowledge transfer. The weights and cluster centres can be directly interpreted as antecedents in a symbolic IF..THEN type rule. The experimental results produced from the algorithm proved that a statistical summary of the weights was too simple to provide accurate rules.

In the case of the iris dataset, the extracted rules were reasonably accurate in describing some of the features of the Setosa class. The other two classes have several features in common which led to an overlap in the intervals assigned to the antecedents. This meant that the extracted rules also had a high degree of overlap and in many cases activated multiple rules. The vibration diagnosis problem, proved to be more difficult. Producing fewer rules gave an overly general solution that had too many overlapping antecedent values which could not resolve the different classes.

Future work intending to overcome these limitations will involve generating a separate rule for each individual RBF unit. This will give a greater accuracy by providing a highly detailed description of the input space. Although such descriptions are accurate they can be difficult to interpret if a large number are generated and have many antecedents. The main disadvantage of using a localist representation scheme is the isolated and fragmented nature of the extracted rules. The problem of rule extraction then becomes one of grouping these rules into appropriate sets.

In conclusion, we argue that the localist character of RBF networks is particulary useful for rule extraction algorithms. Furthermore, there is a trade-off between a large number of extracted rules and a high accuracy versus a small number of rules and a low accuracy. Nevertheless, more research in knowledge extraction is needed for a deeper analysis of RBF networks.

ACKNOWLEDGEMENTS

This work was supported by European funding under the Brite Euram III initiative, project number BE-1313. We would also like to thank Adam Adgar for his advice.

REFERENCES

[1] R. Andrews and S. Geva. Rules and local function networks. In *Rules and Networks-Proceedings of the Rule Extraction From Trained Artificial Neural Networks Workshop, Artificial Intelligence and Simulation of Behaviour*, Brighton UK, 1996.

[2] C. G. Atkeson, A. Moore, and S. Schaal. Locally weighted learning. *Artificial Intelligence Review*, pages 11–73, Feb 1997.

[3] J. Benitez, J. Castro, and J I. Requena. Are artificial neural networks black boxes? *IEEE Transactions on Neural Networks*, 8(5):1156–1164, 1997.

[4] D. S. Broomhead and D. Lowe. Multivariable functional interpolation and adaptive networks. *Complex Systems*, pages 321–355, 1988.

[5] T. Corbett-Clarke and L. Tarassenko. A principled framework and technique for rule extraction from multi-layer perceptrons. In *IEE, Proceedings of the 5th International Conference on Artificial Neural Networks*, pages 233–238, Cambridge, England, July 1997.

[6] R. Fisher. The use of multiple measurements in taxonomic problems. *Annual Eugenics*, 7:179–188, 1936.

[7] J. S. Roger Jang and C. T. Sun. Functional equivalence between radial basis function networks and fuzzy inference systems. *IEEE-NN*, 4(1):156–159, January 1993.

[8] D. Lowe. Characterising complexity in a radial basis function network. In *Proceedings of the 5th International Conference on Artificial Neural Networks*, pages 19–23, Cambridge, UK, 1997.

[9] J. L. McClelland, David E. Rumelhart, and the PDPResearch Group. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 2. MIT Press, Cambridge, MA, 1986.

[10] K. J. McGarry, S. Wermter, and J. MacIntyre. Hybrid neural systems: from simple coupling to fully integrated neural networks. *Neural Computing Surveys*, 2(1), 1999.

[11] J. Moody and C. J. Darken. Fast learning in networks of locally tuned processing units. *Neural Computation*, pages 281–294, 1989.

[12] R. Murray-Smith and T. A. Johansen. Local learning in local model networks. In *IEE Artifical Neural Networks*, pages 40–46, 1995.

[13] J. Murre. The effects of pattern presentation on interference in backpropagation networks. In *The Proceedings of the 14th Annual Conference of the Cognitive Society*, pages 54–59, Hillsdale, NJ, 1992. Lawrence Erlbaum.

[14] C. W. Omlin and C. L. Giles. Extraction and insertion of symbolic information in recurrent neural networks. In V.Honavar and L.Uhr, editors, *Artificial Intelligence and Neural Networks:Steps Towards principled Integration*, pages 271–299. Academic Press, San Diego, 1994.

[15] J. Park and I. W. Sandberg. Universal approximation using radial basis function networks. *Neural Computation*, 3:246–257, 1991.

[16] L. Pratt. *Transfering Previously Learned Back-Propagation Neural Networks to New Learning Tasks*. PhD thesis, Rutgers, State University of New Jersey, May 1993.

[17] M. Roscheisen, R. Hofmann, and V. Tresp. Incorporating prior knowledge into networks of locally-tuned units. In *Computational learning Theory and Natural Learning Systems*, volume III, pages 53–64. 1994.

[18] A. Roy, S. Govil, and R. Miranda. An algorithm to generate radial basis function (rbf)-like nets for classification problems. *Neural Networks*, 8(2):179–201, 1995.

[19] S. Thrun. Extracting rules from artificial neural networks with distributed representations. In G.Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems 7*. MIT Press, San Mateo, CA, 1995.

[20] V. Tresp, J. Hollatz, and S. Ahmad. Representing probabilistic rules with networks of gaussian basis functions. *Machine Learning*, 27:173–200, 1997.

[21] A. Ultsch, R. Mantyk, and G. Halmans. Connectionist knowledge acquisition tool: Conkat. In J. Hand, editor, *Artificial Intelligence Frontiers in Statistics: AI and statistics III*, pages 256–263. Chapman and Hall, 1993.

[22] S. Wermter and R. Sun. *Hybrid Neural Symbolic Systems*. Springer, Heidelberg, 1999 (to appear).

[23] Q. Zhao and Z. Bao. Radar target recognition using a radial basis function neural network. *Neural Networks*, 9(4):709–720, 1996.