Jurnal Teknologi, 45(D) Dis. 2006: 77–96

AN ACTIVE LEARNING APPROACH FOR RADIAL BASIS FUNCTION NEURAL NETWORKS

S. S. ABDULLAH[1] & J. C. ALLWRIGHT[2]

Abstract. This paper presents a new Active Learning algorithm to train Radial Basis Function (RBF) Artificial Neural Networks (ANN) for model reduction problems. The new approach is based on the assumption that the unobserved training data y at input x, lies within a set $F(x) = \{y: \underline{f}(x) \le y \le \overline{f}(x)\}$ where $F(x)$ is known from experience or past simulations. The new approach finds the location of the new sample such that the worst case error between the output of the resulting RBF ANN and the bounds of the unknown data as specified by $F(x)$ is minimized. This paper illustrates the new approach for the case when $x \in R^1$. It was found that it is possible to find a good location for the new data sample by using the suggested approach in certain cases. A comparative study was also done indicating that the new experiment design approach is a good complement to the existing ones such as cross validation design and maximum minimum design.

Key words: Artificial neural networks, radial basis functions, model reduction, active learning, experiment design, metamodeling

Abstrak. Kertas kerja ini membentangkan satu kaedah Pembelajaran Aktif yang baru untuk melatih Jaringan Saraf Buatan (JSB) yang berasaskan Fungsi Asas Jejarian (FAJ) apabila JSB tersebut digunakan untuk menyelesaikan masalah Penurunan Model. Kaedah baru ini berasaskan andaian bahawa data yang diperlukan, y, pada input x, berada dalam sebuah set $F(x) = \{y: \underline{f}(x) \le y \le \overline{f}(x)\}$ di mana $F(x)$ boleh dibentuk menggunakan pengalaman atau pengetahuan awal tentang satu masalah. Kaedah baru ini akan mendapatkan lokasi data baru dengan meminimumkan ralat kes paling buruk antara keluaran JSB dengan had data seperti yang telah ditakrifkan oleh set $F(x)$. Adalah didapati bahawa kaedah yang dicadangkan ini mampu memberikan kedudukan data baru yang baik pada kes-kes tertentu, berbanding dengan data yang diperolehi daripada kaedah sedia ada. Hasil kajian perbandingan antara kaedah yang dicadangkan dengan kaedah yang sedia ada juga disertakan dalam kertas kerja ini yang menunjukkan bahawa kaedah pembelajaran aktif yang dicadangkan merupakan satu penambahan yang baik kepada kaedah pembelajaran aktif yang sedia ada seperti kaedah reka bentuk maksimum minimum atau kaedah *cross validation*.

Kata kunci: Jaringan saraf buatan, fungsi asas jejarian, penurunan model, kaedah pembelajaran aktif, reka bentuk eksperimen, metamodel

[1] Department of Control and Instrumentation Engineering, Faculty of Electrical Engineering Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor. MALAYSIA. Tel: 607-5535455, Fax: 607 - 5566272. Email: Shahrum @fke.utm.my

[2] Control and Power Group, Electrical and Electronic Engineering, Imperial College, South Kensington Campus, London SW72AZ

1.0 INTRODUCTION

An Artificial Neural Network (ANN) needs to 'learn' before it can be effectively used to perform useful tasks. Learning refers to the modification of the ANN parameters so as to bring the map implemented by the ANN as close as possible to a desired map. Three main learning paradigms have emerged: supervised learning, unsupervised learning and reinforcement learning (Bishop, [1]). The focus of this paper will be on Active Learning (a special case of Supervised Learning) for Radial Basis Function (RBF) ANN. In supervised learning, the training examples consist of given inputs and given desired outputs. This means that for every given input pattern, a desired or target output pattern is known and used in the training algorithm.

In this work, the target output pattern is noiseless which is different from typical regression problems where normally there is an assumption of measurement noise corrupting the output. Cases where the output pattern is not corrupted by noise are called model reduction or metamodeling problems [2 - 6]. These problems occur when the training samples are coming from computer simulations in which case no noise is present. Metamodeling involves the determination of simpler models to approximate actual computer simulation models. A simpler model is usually needed when the computational time to simulate the actual model is long and thus it becomes impractical to rely exclusively on simulation for the purpose of design optimization.

Simulations of models involving finite element and fluid dynamics analysis are typical examples of computer programs that require a significant amount of execution time. As an example, a finite element simulation program solving a microwave passive/active circuit problem took about 8 hours execution time on a Pentium-based PC, as reported by Tsai, *et al.* [7]. With the availability of a simpler model, several design issues such as what-if analysis, prediction of a system output, optimization and verification and validation of simulation models can be done using significantly less time since computing the output of an optimized simple model (say a neural network for example) will just be in a matter of minutes, using an equivalent PC. As an example, in a test model reduction for optimization problem in Rashid, *et al.* [8], training a Fuzzy-Neuro simple model on 130 measurement data and the search for its global minimum took only 5.62 minutes of execution time using a similar Pentium-based PC. Hence, although the output of the simple model is only an approximation of the actual output of the complex model, evaluation of this output value is fast and it usually provides enough information especially during the design phase of a project.

In conventional supervised learning, the parameters of an RBF ANN are tuned by minimizing an objective function based on a set of training data. This training paradigm is passive in the sense that the neural network only learns

from the training patterns presented to it by the environment or a teacher. It would be more useful if the ANN could 'suggest' additional training samples itself by using the information on its environment and the existing data samples. Methods of selecting training data from input space have long been studied under the names of Experiment Design [9], Response Surface Methodology [10] and Active Learning (AL) [11]. Here, we will suggest a new AL approach for RBF ANN that takes into consideration the use of important information that is almost always available in real world experiments and that is the range of possible values of the output (or response) of the actual model that one would expect or predict, as a function of the input, based on knowledge and experience.

Among the existing works on active learning for model reduction that have been found, none seem to take into consideration this important factor in the selection of the new samples. Here, we have found that it is possible to utilize this information to find a good location for a new data sample when training an RBF ANN for model reduction or metamodeling.

This paper is organized as follows. Section 2 gives basic terminology and definitions of the RBF ANN. In Section 3, we briefly review existing active learning algorithms. Section 4 formulates the definition of the region of uncertainty within which the output of the actual model is assumed to lie using a set theoretic approach. In Section 5, we present a new active learning algorithm, which uses the definition in 4 to find the location of a new data sample for an RBF ANN. Section 6 is an example showing the performance of the new approach relative to existing deterministic active learning algorithms. Finally, Section 7 concludes this paper.

2.0 RADIAL BASIS FUNCTION NEURAL NETWORKS

The architecture of an RBF ANN is illustrated in Figure 1. The network consists of three layers: an input layer, a hidden layer and an output layer. If the number of output, $Q=1$, the output of the RBF ANN in Figure 1 is calculated according to

$$\eta(x,w) = \sum_{k=1}^{S1} w_{1k}\phi_k\left(\left\|x - c_k\right\|_2\right) \tag{1}$$

where $x \in \Re^{R\times 1}$ is an input vector, $\phi_k(\cdot)$ is a basis function, $\|\cdot\|_2$ denotes the Euclidean norm, w_{1k} are the weights in the output layer, $S1$ is the number of neurons (and centers) in the hidden layer and $c_k \in \Re^{R\times 1}$ are the RBF centers in the input vector space. Equation (1) can also be written as:

$$\eta(x,w) = \phi^T(x)w \tag{2}$$

where

$$\phi^T(x) = \left[\phi_1\left(\|x - c_1\|\right) \;\; \phi_2\left(\|x - c_2\|\right) \;\; \ldots \;\; \phi_{S1}\left(\|x - c_{S1}\|\right)\right] \tag{3}$$

and

$$w^T = \left[w_{11} \;\; w_{12} \;\; \ldots \;\; w_{1S1}\right] \tag{4}$$

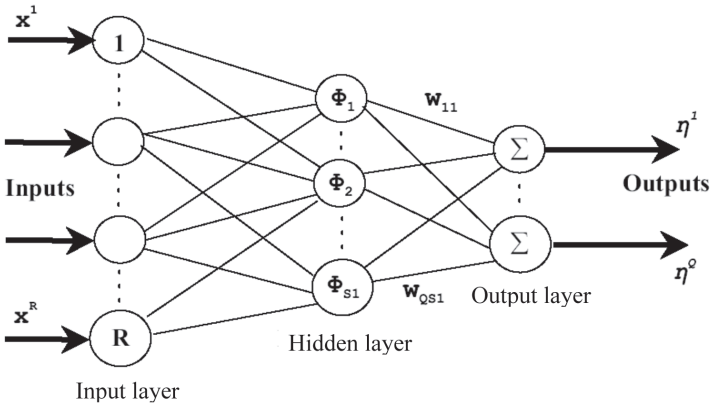

Figure 1 An RBF ANN

The output of the neuron in a hidden layer is a nonlinear function of the distance between its input and the center, c_k. Some typical choices for the functional form of $\phi_k(\cdot)$ are as follows [12]:

$$\phi_k(x) = x \tag{5}$$

$$\phi(x) = x^3 \tag{6}$$

$$\phi(x) = e^{-x^2/\varsigma^2} \tag{7}$$

where the parameter ς controls the "width" of the RBF and is commonly referred to as the spread parameter [12]. The centers c_k are defined points that are assumed to perform an adequate sampling of the input space. Common practice is to select a relatively large number of input vectors as the centers to ensure an adequate input space sampling. After the network has been trained, some of the centers may be removed in a systematic manner without significant degradation of the network mapping performance [12]. Once the centers c_k and the parameter ς have been set, the output layer weights can be found as follows:

(i) Given a set D having N initial input and output training pairs:

$$D = \left\{\left(x_{i_1}, y_{i_1}\right), \left(x_{i_2}, y_{i_2}\right), \ldots, \left(x_{i_N}, y_{i_N}\right)\right\} \tag{8}$$

where and $i_j \in K$ and $K = \{1,2,\ldots,J\}$ with J the number of all possible samples in a discrete input space. Consider the case when the input space is $\Re^{1\times 1}$ and the number of outputs $Q = 1$, equation (1) can be written in a vector matrix form as follows:

$$\begin{bmatrix} \eta(x_{i_1}, w) \\ \vdots \\ \eta(x_{i_N}, w) \end{bmatrix} = \begin{bmatrix} \phi_1(x_{i_1}, c_1) & \cdots & \phi_1(x_{i_1}, c_{S1}) \\ \vdots & \vdots & \vdots \\ \phi_1(x_{i_N}, c_1) & \cdots & \phi_1(x_{i_N}, c_{S1}) \end{bmatrix} \begin{bmatrix} w_{11} \\ \vdots \\ w_{1S1} \end{bmatrix} \tag{9}$$

or

$$\overline{\eta} = \Phi w \tag{10}$$

(ii) A common optimization criterion to use is the quadratic error between the actual and desired ANN outputs

$$E_D(w) = (\boldsymbol{y} - \Phi w)^T (\boldsymbol{y} - \Phi w) \tag{11}$$

where $\boldsymbol{y}$ is the vector of existing output values given by

$$\boldsymbol{y} = \begin{bmatrix} y_{i_1} & y_{i_2} & \cdots & y_{i_N} \end{bmatrix}^T \tag{12}$$

The vector of weights that minimizes (12) can be found [12] and is given by:

$$\hat{w} = (\Phi^T \Phi)^{-1} \Phi^T \boldsymbol{y} = \Phi^{\dagger} \boldsymbol{y}$$

where $\Phi^{\dagger}$ denotes the pseudo-inverse of the nonlinear mapping matrix Φ.

3.0 RELATED WORKS

Existing methods to sequentially select data samples from an environment for non-linear models can be divided roughly into three:
(1) Deterministic methods

Experiment design approaches that fall into this category include the maximum minimum distance (MD) approach [13] and the methods based on Optimum Experiment Design (OED) theory. Examples where methods from optimum

experimental design theory are used to select training data samples for non-linear models can be found in [9, 11, 14, 15]. In Cohn [11], the prediction variance equation of a neural network was calculated, while Choueiki [14] uses the prediction variance of a second order polynomial, instead of that of the neural networks, to guide in sampling new data.

Mackay [15] looks at the sampling problem from the Bayesian perspective. Using Taylor expansion of a neural network's output, he derives the equation of the output prediction variance. He also proved that if the additional training data point is taken where this variance is largest, then the rule will be the same as that using the D-optimal criterion.

Similarly in Atkinson, *et al.* [9] the D-optimal design was extended to non-linear models by linearizing the model by Taylor series expansion and using the linearized model to compute the standardized variance (SV) equation. The steps that they have suggested for OED when the model is non-linear are as follows:

(i) Start with a preliminary estimate
(ii) Linearize the model by Taylor expansion
(iii) The optimum locations of the new data samples for the linearized model are determined and new measurements are obtained at the optimum locations
(iv) Analyze (iii). If sufficient, stop. Otherwise repeat step (ii) until sufficient accuracy is achieved

(2) Probabilistic or stochastic method

Methods belonging into this category are gradient free methods that rely largely on random search where the new samples are obtained from either a certain probability distribution function [16] or using genetic algorithms to generate the location of the samples [17].

(3) Neural Networks Optimized on Different Data Sets

This involves training various neural network models on different sets or pseudo replicates of the original training data samples. A measure of discrepancy or ambiguity of the different optimized networks is defined and used to obtain new training samples from the input space. These methods are also called "cross-validation" (CV) (Jin, *et al.* [2]) or "bootstrapping" by some researchers.

In this paper, only the deterministic methods and the cross-validation approach of active data selection will be compared with the new AL approach suggested in Section 5.

4.0 REPRESENTING THE UNCERTAINTY OF THE UNOBSERVED DATA

In this section, the uncertainty in the knowledge of the complex function is defined to facilitate the active data selection process. The function to be approximated by the neural network has the form $f:\Re^n \to \Re^m$ and we are interested in approximating on a domain χ of $\Re^n$. Experiment design are facilitated by defining a set-valued map $f:\Re^n \to \Re^m$ such that $f(x) \in F(x)$ for all $x \in \chi$. The set $F(x)$ defines the uncertainty in the knowledge of the value of $f(x)$. We usually suppose that $F(x)$ has the form:

$$F(x) = \left\{ y: \underline{f}(x) \le y \le \overline{f}(x) \right\}$$

for appropriate $\underline{f}, \overline{f}: R^n \to R^m$ where $\underline{f}(x) \le y \le \overline{f}(x)$ means

$$\underline{f}(x)_i \le y_i \le \overline{f}(x)_i, \quad i = 1 \ldots m$$

Suppose initially, before carrying out any experimental measurements, that we make the guesstimate $f(x) \in F_\phi(x)\ \forall\ x \in \chi$, where the symbol ϕ indicates that no experimental results have been used. And we assumed that F_ϕ has the form

$$F_\phi(x) = \left\{ y: \underline{f_\phi}(x) \le y \le \overline{f_\phi}(x) \right\} \forall x \in \chi$$

for appropriately specified $\underline{f_\phi}, \overline{f_\phi}: \Re^n \to \Re^m$. A basic problem is that of modifying F_ϕ to take account of any experimental measurements that have been made to reduce the amount of uncertainty in F_ϕ. A set-theoretic approach is developed next.

Assume $f(x)$ has the uniform k-Lipschitz property that for any finite $k > 0$ there exists an $\in_k > 0$ so that for all $\bar{x} \in \chi$

$$\left\| f(x) - f(\bar{x}) \right\| \le k \left\| x - \bar{x} \right\|, \forall x \in \bar{x} + \frac{\in_k}{2} B$$

where $B = \left\{ x: \|x\| \le 1 \right\}$ and $x \in \bar{x} + \frac{\in_k}{2} B$. We shall use this idea by assuming that, using prior knowledge, a k and $\in_k$ can be chosen such that for any input value x near a past measurement, the value of $f(x)$ lies within the cone shown in Figure 2. Taking each data samples x_i $(i = 1 : N)$ to be an, this yields

$$\left\| f(x) - f(x_i) \right\| \le k \left\| x - x_i \right\| \tag{13}$$

for all x such that $\|x - x_i\| < \frac{\in_k}{2}$. As a refinement of this, a different k and $\in_k$ can be used for different parts of the input space but this will cause further complications, which will be avoided here.

This formulation provides information regarding the function f that can be used to modify the initial set valued map $F_\phi(x)$. We shall treat the simple scalar case first. For this case, condition (13) implies that $f(x) \in L_i$ for all x such that $\|x - x_i\| < \frac{\in_k}{2}$, where

$$L_i = \left\{ y : \|y - f(x_i)\| \le k\|x - x_i\| \right\} \qquad i = 1{:}N \tag{14}$$

L_i is illustrated with the shaded area in Figure 2.
If we have more than one existing data samples, define

$$L = \cup_i L_i \qquad i = 1{:}N \tag{15}$$

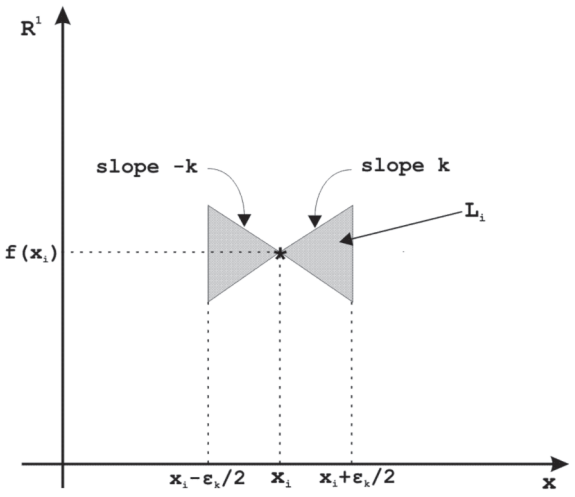

Figure 2 L_i showing the initial set value map when N=1

This is illustrated by the grey shaded areas in Figure 3. Next, define to be the points $(y, x) \in R^m \times R^n$ that have $x \in x_i + \frac{\varepsilon_k}{2} B$ so

$$L_i^C = \left\{ \begin{pmatrix} y \\ x \end{pmatrix} : \|y - f(x_i)\| > k\|x - x_i\|, \|x - x_i\| < \frac{\varepsilon_k}{2} B \right\} \tag{16}$$

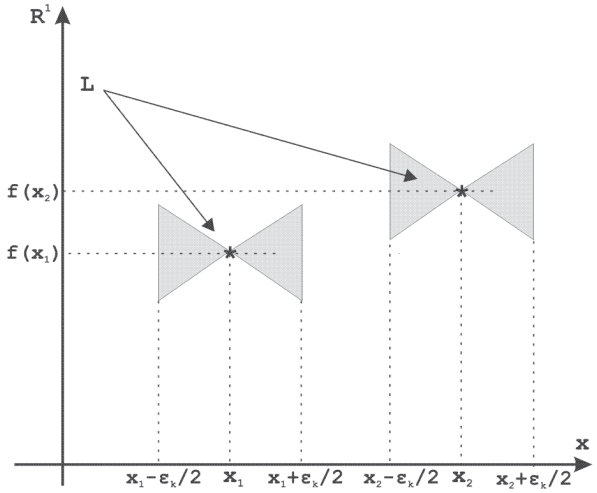

Figure 3 *L* showing the initial set value map if *N>1*

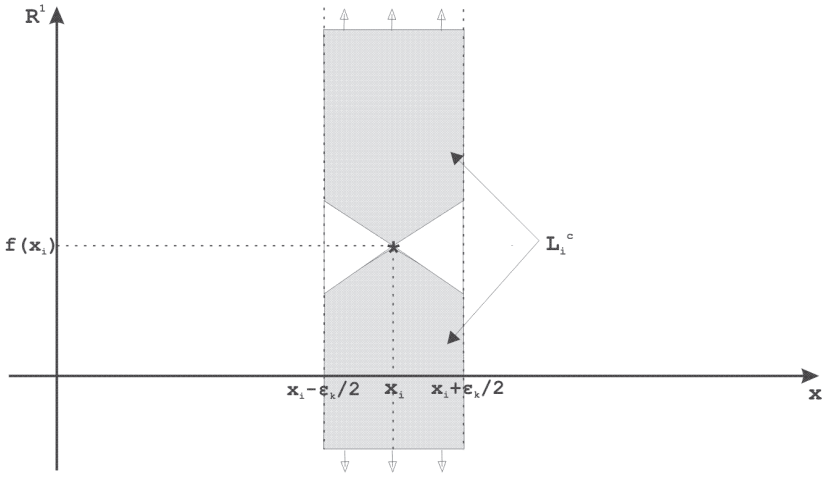

Figure 4 L_i^c is the region where the function should not be in for the case when *N=1*

where $i = 1 : N$. This is illustrated by the shaded area in Figure 4. Consequently, for several data samples, define

$$L^c = \bigcup_i L_i^c \tag{17}$$

The shaded areas in Figure 5 illustrate L^c. Suppose D is the set of measured experimental data in that

$$D = \{(x_i, f(x_i)),\ i = 1 : N\}$$

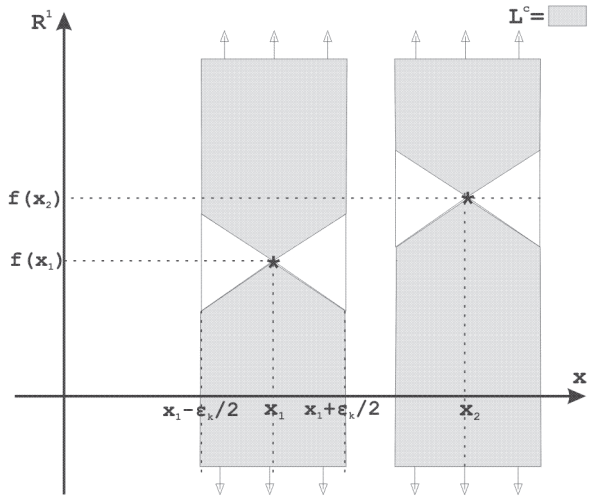

Figure 5 L_c (shaded area) is the region where $f(x)$ should not be in when $N>1$.

Then we shall next shown how to take account of D to generate $F_D(x)$ with (ideally) $FD(x) \subset F\phi\ (x)$ indicating that there is less uncertainty. This will be done in a set theoretic way.

Now

$$\text{Graph } (F_\phi) = \{(x, y) \in \chi \times R^m : y \in F_\phi\ (x)\}$$

This is illustrated in Figure 6. We shall try to use the data to generate a set F_D with a "smaller graph". We will choose

$$F_D = (F_\phi - F_\phi \cap L^c) \cup L \tag{18}$$

as the updated graph. Here, $(F_\phi - F_\phi \cap L^c)$ removes the part of F_ϕ that is to be replaced by all the L_i's and so $F_D = (F_\phi - F_\phi \cap L^c) \cup L$ yields the updated graph. If each L_i is consistent with F_ϕ in that $L_i \subset F_\phi\ \forall\ i$, then $L \subset F_\phi$ and consequently F_D is 'smaller than' F_ϕ since its "area" is reduced. Figure 7 is an example to illustrate the updated graph F_D for the case when $L \subset F_\phi$ or $L \cap F_\phi = L$ (i.e. when all region of L lies within F_ϕ). The initial set F_ϕ is as illustrated in Figure 6. Since F_D has the form

$$F_D(x) = \left\{y: \underline{f_D}(x) \le y \le \overline{f_D}(x)\right\} \forall\, x \in \chi$$

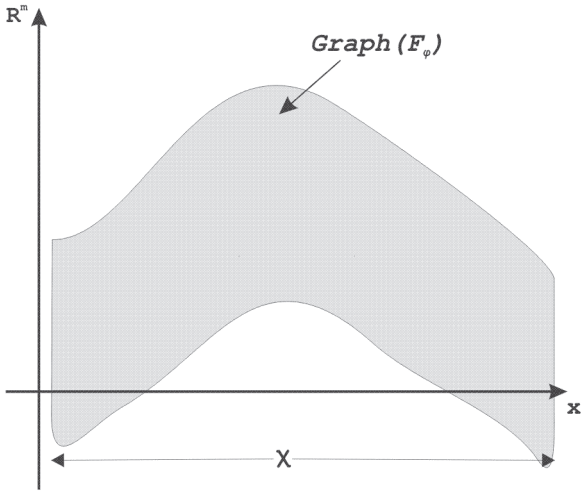

Figure 6 F_ϕ representing the initial set value map.

then in this case, $\underline{f_D}(x)$ is indicated by the – line while $\overline{f_D}(x)$ is indicated by the -.- line in Figure 7. From this figure, it can be observed that $F_D \subset F_\phi$ in this case and the "area" of F_ϕ has been reduced, reflecting the fact that we are more "sure" of *f(x)* after obtaining the data samples. Note that formula (18) also works for other cases such as when $(L \cap F_\phi) \neq L$ (i.e. when some region of L lies outside F_ϕ) and $\cap_i L_i \neq \varnothing$. For details, refer to [6].

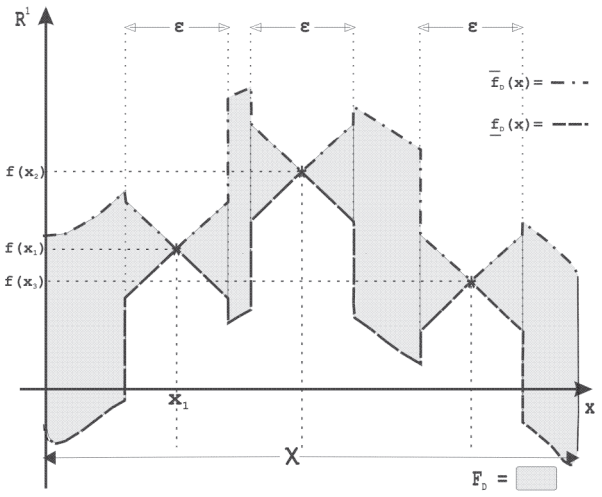

Figure 7 Example, $F_D = (F_\phi - F_\phi \cap L^c) \cup L$

5.0 A NEW ACTIVE LEARNING ALGORITHM FOR RBF ANN

Here we propose a new sequential Active Learning approach for RBF ANN utilizing the bounds of the unknown data. We will call the approach, the Worst Case (WC) approach. Let the new data point to be added to the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ be $\left(x_i, y_i\right)$. Since the centers c_k of the RBF ANN are set to be equal to the value of the available input data, i.e.

$$c_1 = x_{i_1}, c_2 = x_{i_2}, \ldots, c_{S1} = x_{i_N} \tag{19}$$

then the arrival of $\left(x_i, y_i\right)$ implies a new center $\tilde{c} = x_i$ will be added to the existing ones. To facilitate the derivation, define the matrix $\tilde{\Phi}$ as the matrix that takes into account the additional new data sample and the new center:

$$\tilde{\Phi} = \begin{bmatrix} \phi_1\left(x_{i_1}, c_1\right) & \ldots & \phi_{S1}\left(x_{i_1}, c_{S1}\right) & \phi_i\left(x_{i_1}, \tilde{c}\right) \\ \vdots & \vdots & \vdots & \vdots \\ \phi_1\left(x_{i_N}, c_1\right) & \ldots & \phi_{S1}\left(x_{i_N}, c_{S1}\right) & \phi_i\left(x_{i_N}, \tilde{c}\right) \\ \phi_1\left(x_{i_1}, c_1\right) & \ldots & \phi_{S1}\left(x_i, c_{S1}\right) & \phi_i\left(x_i, \tilde{c}\right) \end{bmatrix} \tag{20}$$

where $\phi(x)$ can be any of the radial basis functions (5)-(7). Also define

$$\tilde{\phi}(x)^T = \left[\phi_1\left(\|x - c_1\|\right) \;\; \phi_2\left(\|x - c_2\|\right) \;\; \ldots \;\; \phi_{S1}\left(\|x - c_{S1}\|\right) \;\; \phi\left(\|x - \tilde{c}\|\right)\right] \tag{21}$$

and

$$\tilde{w} = \begin{bmatrix} w_{11} & w_{12} & \ldots & w_{1S1} & \tilde{w}_1 \end{bmatrix} \tag{22}$$

Next, the errors when additional data at $\left(x_i, y_i\right)$ is added to the existing data set D for the RBF ANN are defined as

$$E_{\tilde{D}}(\tilde{w})_p \begin{cases} = \left|y_i - \tilde{\phi}(x_i)^T \tilde{w}\right| + \sum_{j=1}^{N} \left|y_{i_j} - \tilde{\phi}(x_{i_j})^T \tilde{w}\right| & \text{if } p = 1 \\ = \left(y_i - \tilde{\phi}(x_i)^T \tilde{w}\right)^2 + \sum_{j=1}^{N} \left(y_{i_j} - \tilde{\phi}(x_{i_j})^T \tilde{w}\right) & \text{if } p = 2 \\ = \max\limits_{j \in [1,\ldots N]} \begin{vmatrix} y_{i_j} - \tilde{\phi}(x_{i_j})^T \tilde{w} \\ y_{i_j} - \tilde{\phi}(x_i)^T \tilde{w} \end{vmatrix} & \text{if } p = \infty \end{cases} \tag{23}$$

Hence the weights $\tilde{w}_{x_i}$ that minimize $E_{\tilde{D}}(\tilde{w})_p$ are found by

$$\tilde{w}_{x_i} \in \arg \min_{\tilde{w} \in \Re^{N+1}} \max_{y_i \in \left[\underline{f_D}(x_i), \overline{f_D}(x_i)\right]} E_{\tilde{D}}(\tilde{w})_p \tag{24}$$

Therefore the new data will be obtained at location x_i that has the minimum worst case error which can be defined as

$$x_i^* \in \arg\min_{x_i \in \chi} \overline{E}(\tilde{w}_{x_i})_q \tag{25}$$

where $\overline{E}(\tilde{w}_{x_i})_q$ are worst case errors defined by

$$\overline{E}(\tilde{w}_{x_i})_q = \max_{\substack{f(x_1) \in \left[\underline{f_D}(x_1), \overline{f_D}(x_1)\right] \\ \vdots \\ f(x_J) \in \left[\underline{f_D}(x_J), \overline{f_D}(x_J)\right]}} \sum_{i=1}^{J} \left| f(x_i) - \tilde{\phi}(x_i)^T \tilde{w}_{x_i} \right| \qquad \text{if } q = 1 \tag{26}$$

$$= \max_{\substack{f(x_1) \in \left[\underline{f_D}(x_1), \overline{f_D}(x_1)\right] \\ \vdots \\ f(x_J) \in \left[\underline{f_D}(x_J), \overline{f_D}(x_J)\right]}} \sum_{i=1}^{J} \left(f(x_i) - \tilde{\phi}(x_i)^T \tilde{w}_{x_i} \right)^2 \qquad \text{if } q = 2 \tag{27}$$

$$= \max_{\substack{f(x_1) \in \left[\underline{f_D}(x_1), \overline{f_D}(x_1)\right] \\ \vdots \\ f(x_J) \in \left[\underline{f_D}(x_J), \overline{f_D}(x_J)\right]}} \left| f(x_i) - \tilde{\phi}(x_i)^T \tilde{w}_{x_i} \right| \qquad \text{if } q = \infty \tag{28}$$

The WC AL algorithm can then be summarized as follows:

(i) Define the bounds F_ϕ using existing knowledge of the unknown function *f(x)*.

(ii) Use existing samples $\boldsymbol{x}$ and $\boldsymbol{y}$ to form F_D.

(iii) Obtain the location of the new sample x_i^* using Equation (25).

(iv) Obtain a new measurement y_i (i.e. $f(x_i^*)$).

(v) Update F_D using the existing samples and the new sample $\left(x_i^*, f(x_i^*)\right)$.

(vi) Repeat steps (iii) - (v) until the desired amount of new data has been obtained.

6.0 CASE STUDY

Consider the case when $f(x)$ is given by

$$f(x) = e^{x}\cos(2\pi x)$$

The initial input vector x is given by

$$\boldsymbol{x} = [\,-3.0 \quad -1.5 \quad 0 \quad 1.5 \quad 3.0\,]^T \tag{29}$$

and the initial output measurements are

$$y = [\,0.05 \quad -0.22 \quad 1.00 \quad -4.48 \quad 20.09\,]^T \tag{30}$$

Define the error, $E(w)_\infty$, which can be computed assuming $f(x)$ known as:

$$E(w)_\infty = \max\left| f(x_i) - \tilde{\phi}(x_i)^T \tilde{w} \right| \tag{31}$$

$\eta(x, \omega)$ is the RBF ANN as defined in equation (1) with $\phi_k(x) = e^{-x^2/\varsigma^2}$. The parameter ς for the RBF ANN was set to 1.0. The centers are set equal to the available input samples as given in Equation (19). The input space is given by $\chi = [-3,3]$. The optimized η, the data samples, $\underline{f_D}(x)$, $\overline{f_D}(x)$ and the location for the new sample that minimized $E(w)_\infty$ and $E(w)_2$ (assuming $f(x)$ known) are shown in Figure 8. The details of the colours and symbols used in Figure 8 are provided in Table 1. The initial bounds where formed assuming the exponential property of $f(x)$ is known and ensuring that $f(x) \in F_\phi$. Cases when $f(x) \notin F_\phi$ will not be covered here.

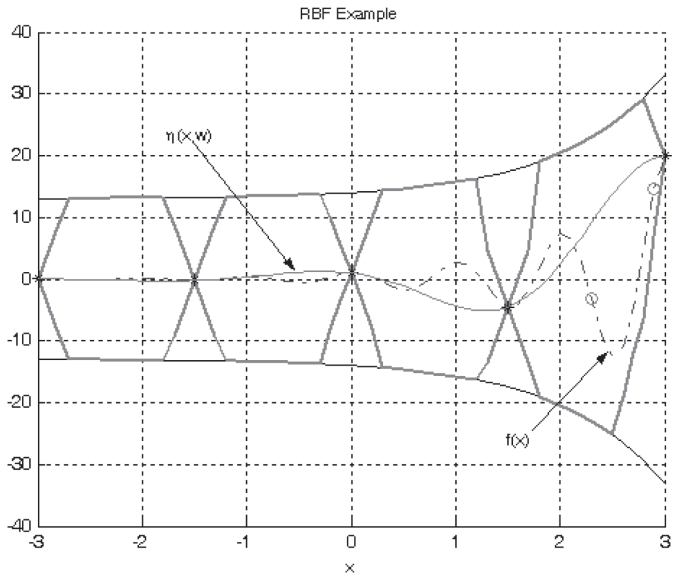

Figure 8 A Model Reduction Problem

Table 1 Symbols used in Figure 8 describing the problem

Quantity	Symbol
$\overline{f}_{\phi}$ and $\underline{f_{\phi}}$	Black solid lines
$\overline{f}_{D}$ and $\underline{f}_{D}$	Grey solid lines
$\eta(x, \omega)$	Labeled solid line
$f(x)$	Dash-dot line
Initial x_i	Black asterisks

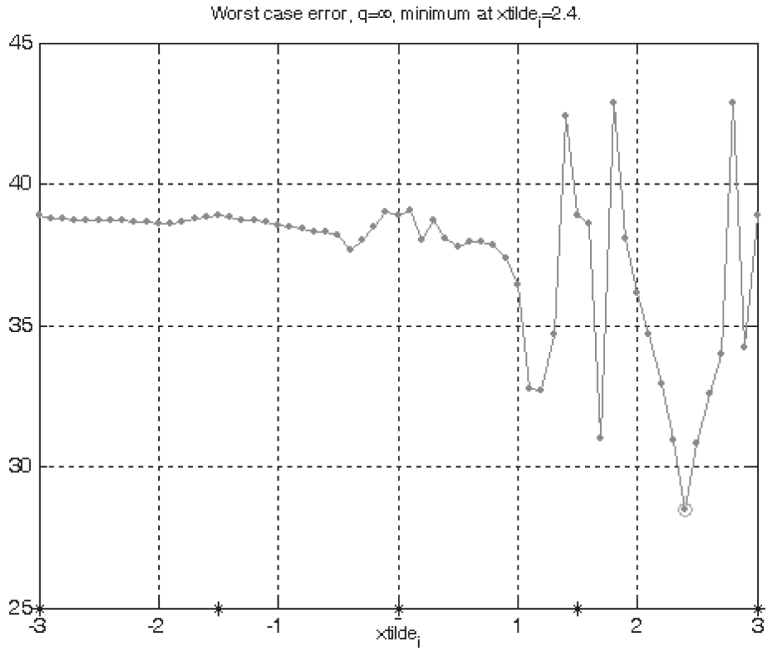

Figure 9 $E(\tilde{w}_{x_i})_{\infty}$(Worst case error)

$E(w)_{\infty}$ in this case is shown in Figure 10 showing a minimum at $x = 2.3$ which is close to the location suggested by the WC approach. This defines the location of the new data that if sampled will cause the greatest reduction of error $(E(w)_{\infty})$. Other Active Learning approaches reviewed in Section 3 were also used to find the new location. The results are summarized in Table 2. Here only the CV approach manages to find the optimum location for the new sample (at $x = 2.3$).

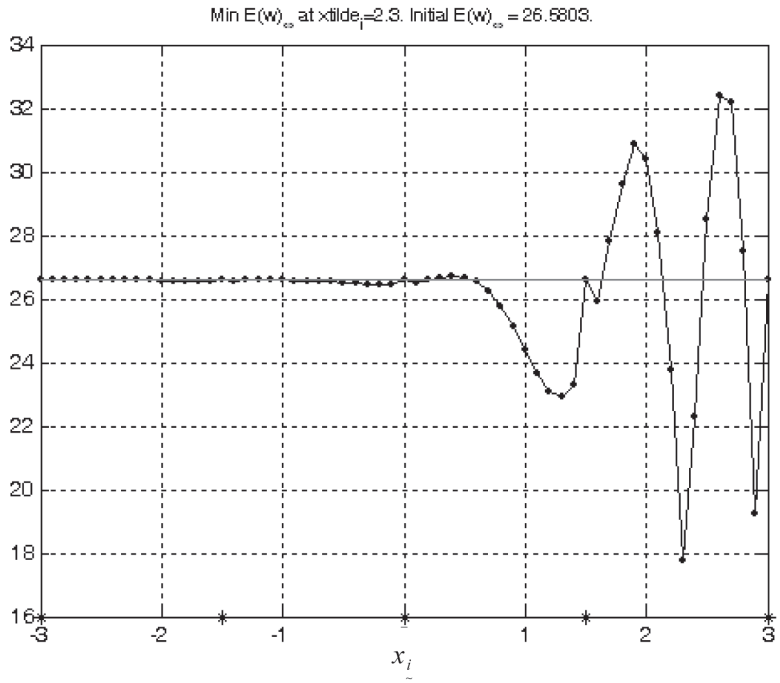

Figure 10 (Actual error)

Table 2 $E(\omega)_\infty$ for different experiment design approaches

	Methods of Experiment Design			
	WC	MD	CV	SV
Location of new x_i =	2.4	-2.3, -0.8 0.8 or **2.3**	**2.3**	Initial x_i
$E(\omega)_\infty$ after sampling the new point	22.3	26.6 or **17.8**	**17.8**	26.6
Reduction of $E(\omega)$(= (Initial $E(\omega)_\infty = 26.6$)	4.3	0 or **8.8**	**8.8**	0

The experiment was repeated 15 times using different locations for the initial data samples. The error $E(w)_\infty$ was calculated after obtaining the new sample using the different experiment design methods. The results were recorded in Table 3 and summarized using the box diagram in Figure 11. The cases when the WC approach perform better than the other AL methods are highlighted using bold numerals in the table. The box diagram shows the highest, the lowest, the median, the upper quartile and the lower quartile $E(w)_\infty$ values (see Figure

Table 3 $E(\omega)_\infty$ after sampling the new point (the lower the better)

	$E(\omega)_\infty$ after sampling the new point			
Initial x_i locations	WC	MD	CV	SV
-3.0, -0.8, 0.8, 2.3, 3.0	20.6	18.3	17.3	26.7
-3.0, -2.3, 0.3, 2.6, 3.0	31.6	39.3	25.7	45.4
-3.0, -0.8, 0.0, 2.0, 3.0	**20.3**	**28.0**	**39.8**	**28.0**
-3.0, -1.7, 0.6, 0.6, 3.0	21.2	27.1	18.4	28.0
-3.0, -2.6, 1.7, 1.9, 3.0	47.2	33.4	35.0	37.9
-3.0, -1.3, 0.6, 2.5, 3.0	51.0	34.8	22.6	43.9
-3.0, -0.8, 1.1, 2.5, 3.0	36.8	31.7	16.7	31.4
-3.0, -2.7, 0.1, 1.3, 3.0	19.3	27.7	17.8	27.7
-3.0, -1.7, 1.7, 1.7, 3.0	**14.7**	**26.4**	**18.2**	**26.4**
-3.0, -1.4, 2.6, 2.9, 3.0	46.1	59.3	43.0	46.1
-3.0, -1.3, 0.5, 1.2, 3.0	17.8	28.2	17.6	28.8
-3.0, -1.4, 0.5, 0.8, 3.0	19.0	27.1	17.9	25.8
-3.0, -2.0, 0, 2.0, 3.0	**20.3**	**28.0**	**39.9**	**28.0**
-3.0, -1.5, 0, 1.5, 3.0	22.3	26.6	17.8	26.6
-3.0, -1.8, 1.1, 2.0, 3.0	**15.1**	**27.7**	**25.5**	**27.6**

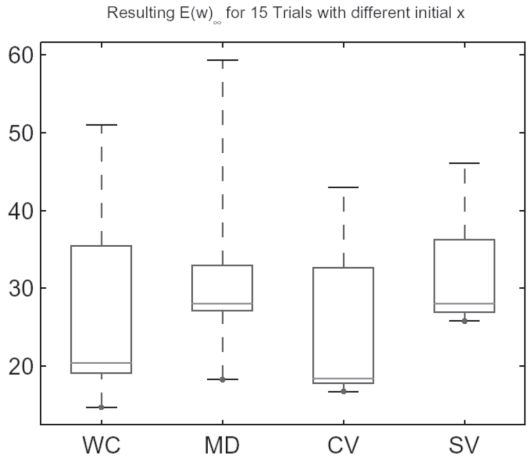

Figure 11 Box diagram of Table 3.

12). The median $E(w)_\infty$ for the 15 trials was lowest using the CV approach and the highest using the MD approach. For this RBF example, the approach that has the lowest median is the CV approach, while the SV approach have the lowest deviation from the median for the 15 trials and the WC approach has the lowest error for a particular trial.

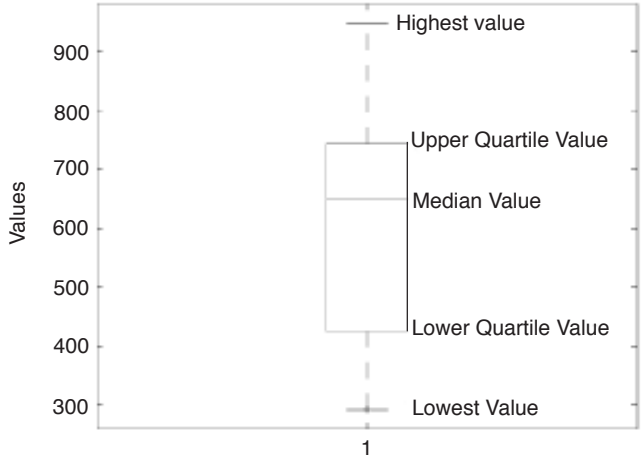

Figure 12 Descriptions of a box diagram

This is a good example to show that there is no particular sequential experiment design approach that is the best for all types of problem. The performance of the experiment design approaches are dependent on several factors which include the actual f, the simple model and the initial location of the data samples x_i. However this example demonstrated that it is possible to use the suggested WC approach to find a good location for a new data sample for model reduction using RBF ANN and hence the approach is a good complement to the existing ones.

7.0 CONCLUSIONS

This work suggests a new AL algorithm for RBF ANN used in a metamodeling problem. The new approach takes into consideration the actual physical constraints of the problem. Among the four methods investigated, only the WC approach takes into consideration the actual physical information available of the unknown model. Hence this method is a good complement to the other approaches where the estimates are based only on the existing data distribution and the structure of the simple model η. Also in certain cases as shown in the case study, the WC approach managed to outperform the existing methods. This indicates the advantage of the proposed approach on certain types of problems. However, further investigations have to be conducted to identify the type of problems that are most suitable for the WC approach.

ACKNOWLEDGEMENTS

The authors would like to express their gratitude to Universiti Teknologi Malaysia for providing the financial assistance throughout the duration of this project.

AUTHORS

Shahrum Shah Abdullah is a lecturer at the Universiti Teknologi Malaysia. His research interests include Experiment Design and the use of Artificial Intelligence to solve control and optimization problems.

John Allwright is currently a senior lecturer at the Imperial College of Science, Technology and Medicine, UK. His research interests include optimisation problems for symmetric matrices, Neural networks for control and Orthogonal Lyapunov transformations.

REFERENCES

[1] Bishop, M. 1995. *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press.

[2] Jin, R., W. Chen, and A. Sudjianto. 2002. On Sequential Sampling for Global Metamodeling in Engineering Design. Proceedings of Design Engineering Technical Conference and Computers and Information in Engineering Conference. Montreal, Canada.

[3] Martin, J. D., and T. W. Simpson. 2002. Use of Adaptive Metamodeling for Design Optimization. 9th AIAA/ISSMO Symposium on Multidisciplinary Analysis and Optimization. Atlanta, Georgia.

[4] Johnson, K. E., K. W. Bauer Jr., J. T. Moore, and M. Grant. 1996. Metamodeling Techniques in Multidimensional Optimality Analysis for Linear Programming. Math. Comput. Modeling. 23(5): 45-60.

[5] Kleijnen, J. P. C. 1987. *Statistical Tools for Simulation Practitioners*. New York: Marcel Dekker.

[6] Abdullah, S.S. 2003. Experiment Design for Deterministic Model Reduction and Neural Network Training. *Ph.D. Thesis*. Imperial College of Science, Technology and Medicine, Department of Electrical and Electronics Engineering. London, U.K.

[7] Tsai, H. P., Y. Wang, and T. Itoh. 2002. An Unconditionally Stable Extended (USE) Finite-Element Time-Domain Solution of Active Nonlinear Microwave Circuits Using Perfectly Matched Layers. *IEEE Transactions on Microwave Theory and Techniques*. 50 (10): 2226-2232.

[8] Rashid, K., M. Farina, J. A. Ramirez, J. K. Sykulski, and E. M. Freeman. 2001. A Comparison of Two Generalized Response Surface Methods for Optimization in Electromagnetics. COMPEL: *The International Journal for Computation and Mathematics in Electrical and Electronic Engineering*. 20(3): 740-752.

[9] Atkinson, A. C., and A. N. Donev. 1992. *Optimum Experimental Design*. Oxford: Oxford Science Publications.

[10] Myers, R. H., and D. C. Montgomery. 1995. *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. New York: Wiley.

[11] Cohn, D.A. 1994. Neural Network Exploration Using Optimal Experiment Design. Advances in Neural Information Processing System. 6: 679 - 686.

[12] Ham, F. M., and I. Kostanic. 2001. *Principles of Neurocomputing for Science and Engineering. Singapore:* McGraw-Hill,.

[13] Johnson, M., L. Moore, and D. Ylvisaker. 1990. Minimax and Maximin Distance Designs. *Journal of Statistical Planning and Inference.* 26: 131 - 148.

[14] Choueiki, M.H. 1999. Training Data Development with the D-Optimality Criterion. *IEEE Transactions on Neural Networks.* 10(1).

[15] Mackay, D. J. C. 1992. Information-based Objective Functions for Active Data Selection: Neural Computation. 4: 590 - 604.

[16] Fukumizu, K. 2000. Statistical Active Learning in Multilayer Perceptrons, *IEEE Transactions on Neural Networks.* 11(1): 129 - 144.

[17] Zhang, B.T., and G. Veenker. 1991. Neural Networks that Teach Themselves through Genetic Discovery of Novel Examples. *IEEE International Joint Conference on Neural Networks.* 690 - 695.

[18] Kindermann, J., G. Paass, and F. Weber. 1995. Query Construction for Neural Network Using the Bootstrap. *Proc. Int. Conf. Artificial Neural Networks.* 95: 135 - 140.

[19] Raychaudhuri, T., and L. G. C. Hamey. 1995. Minimisation of Data Collection by Active Learning. Proceedings, *IEEE International Conference on Neural Networks.* 3: 1338-1341.