Monograph:
Lim, Chee Peng and Harrison, R.F. (1997) A Multiple Neural Network Architecture for Sequential Evidence Aggregation and Incomplete Data Classification. Research Report. ACSE Research Report 684 . Department of Automatic Control and Systems Engineering

A Multiple Neural Network Architecture for Sequential Evidence Aggregation and Incomplete Data Classification

Chee Peng Lim
School of Industrial Technology
Universiti Sains Malaysia
11800 Minden, Penang
Malaysia

Robert F. Harrison
Department of Automatic Control and Systems Engineering
University of Sheffield
Mappin Street, Sheffield S1 3JD
United Kingdom

Research Report No 684

August 1997

ABSTRACT

In this paper, a multiple neural network architecture is proposed for undertaking the problems associated with incomplete or missing data in on-line learning and classification tasks. An autonomously learning neural network classifier, which has been previously devised based upon the integration of Fuzzy ARTMAP and the Probabilistic Neural Network, is employed as the basis for the development of the multiple neural network system. Each classifier is dedicated to handle a set of input features independently, and produces a prediction of the target class. Bayes' theorem is then applied to combine the outcomes from disparate classifier modules sequentially. Applicability of the multiple neural network system is demonstrated using a simulated data set and a real medical diagnosis database, and the results are compared with other approaches.

Keywords: multiple neural networks, on-line learning, pattern classification, incomplete data, Bayes' theorem, medical diagnosis

1 INTRODUCTION

In many real applications, data is sparse and of variable quality. Often, a complete set of input features may not be available for immediate use. This problem occurs in many situations where data is drawn from more than one source or by different techniques. One of the limitations of many neural network models is that no provision is made to handle incomplete data sets or missing data. It is assumed that all the input items are accessible for the network to generate a result. However, it is not unusual to encounter the missing data scenario in many "real-world" applications. For example, in the case of medical diagnosis, some of the data items such as ECG measurements, X-ray, and other radiographic images need to be interpreted and encoded by domain experts and may not be available instantaneously.

In our previous work [1], we have developed a hybrid network which is capable of incremental learning, and thus avoid the problems of catastrophic forgetting and re-training when operating on-line in non-stationary environments. The network is based upon an integration of two network architectures: Fuzzy ARTMAP [2] and the Probabilistic Neural Network [3]. This hybrid network has been shown to be capable of providing outputs which estimate the Bayesian *a posteriori* probabilities, and of achieving the Bayes optimal results autonomously without prior knowledge of impending changes in data the environment. It also achieves comparable performance with other approaches in a number of benchmark problems [1, 4], but with the ability of on-going (causal) learning.

Based on the hybrid network described in [1], a multiple neural network architecture is proposed here for incremental learning and classification of incomplete data sets. The system makes use of Bayes' theorem to combine decisions from multiple classifier modules sequentially. When given an incomplete data set, the important feature items can be grouped together and presented to a classifier to give an initial prediction. Then, data collected later can be fused to another classifier to reinforce or counteract the initial predictions. As a result, the multiple classifier system is able to make use of more and more information in generating a predicted output with more confidence as time goes on.

2 A HYBRID NETWORK

2.1 Fuzzy ARTMAP

It is well documented that the family of Adaptive Resonance Theory (ART) networks offer an alternative for solving the so-called stability-plasticity dilemma—how a learning system can absorb new information without forgetting previously learned information [5]. More recently, a supervised ART network known as Fuzzy ARTMAP (FAM) which realises a synthesis of ART and fuzzy logic has been introduced. Figure 1 depicts a schematic diagram of the FAM network. It consists of two Fuzzy ART [6] modules, ART_a and ART_b, linked by a map field, F_{ab}. The ART_a (ART_b)

module has two layers of nodes: F_{1a} (F_{1b}) is the input layer; and F_{2a} (F_{2b}) is a dynamic layer where each node encodes a prototype pattern of a cluster of input patterns, and the number of nodes can be increased when necessary.

The key feature of FAM is in the inclusion of a novelty detector in ART_a to measure against a threshold the similarity between the prototype patterns stored in the network and the input pattern. When the match criterion is not satisfied, a new node is created, and the input is coded as its prototype pattern. As a result, the number of nodes grows with time, subject to the novelty criterion, in an attempt to learn a good network configuration autonomously and on-line. As different tasks demand different network structures, this learning approach thus avoids the need to specify a pre-defined static network size, or to re-train the network off-line.

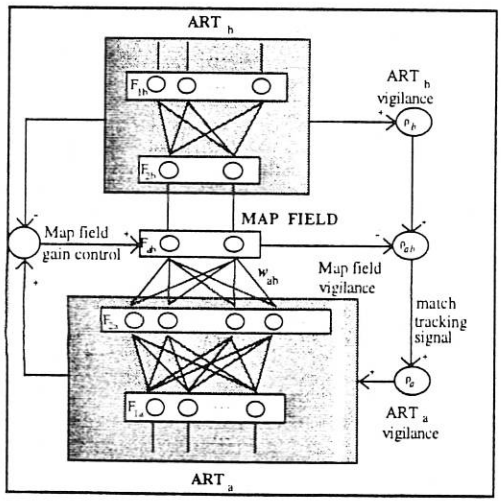

Figure 1 The Fuzzy ARTMAP network

During supervised learning, ART_a receives a stream of input vectors, $\{\boldsymbol{A}\}$, whereas ART_b receives the corresponding target-class vectors, $\{\boldsymbol{B}\}$. In general, ART_b consists of an independent Fuzzy ART module to self-organise the target vectors. However, in one-from-N classification (*i.e.*, each input pattern belongs to only one of the N possible output classes), ART_b can be replaced by a single layer containing N nodes. Then, the N-bit teaching signal can be coded to have unit value corresponding to the target category and zero for all others.

The learning algorithm of FAM is similar to the sequential leader clustering algorithm [7]. However, FAM does not directly associate input patterns at ART_a with target patterns at ART_b. Rather, input patterns are first classified into prototypical category clusters before being linked with their target outputs via a map field. At each input pattern presentation, this map field establishes a link from the winning category prototype in F_{2a} to the target output in F_{2b}. This association is used, during testing, to recall a prediction when an input is presented to ART_a.

2.2 The Probabilistic Neural Network

The Probabilistic Neural Network (PNN) is a neural network model that implements the Bayes' theorem in its learning methodology. It learns instantaneously in one-pass through the data samples and is able to form complex decision boundaries which approximate asymptotically the Bayes optimal limits. In addition, the decision boundaries can be modified on-line when new data is available without having to re-train the network. Another advantage of the PNN is its speed of learning, which is often orders of magnitude faster than that of the Multi-Layer Perceptron (MLP) trained with back-propagation [3].

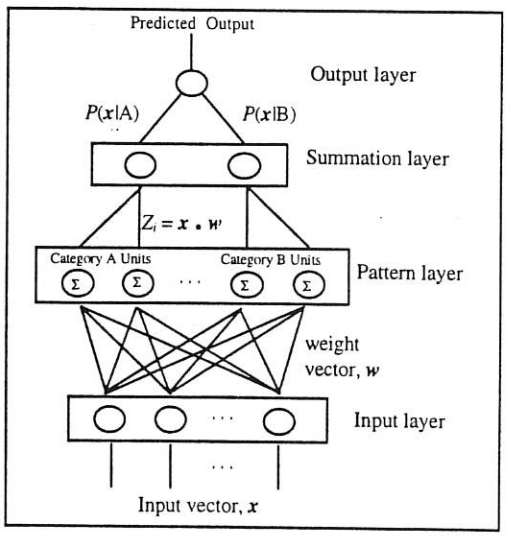

Figure 2 The Probabilistic Neural Network

The key feature of the PNN is its ability to estimate the probability density functions (pdfs) based on the data samples by using the Parzen-windows technique [8]. Figure 2 depicts a schematic diagram of the PNN for binary classification tasks (class A or B). The PNN consists of four layers of nodes: the input layer, pattern layer, summation layer, and output layer. Nodes in the pattern layer are organised in groups corresponding to different target classes. The pattern nodes belonging to the same output are then linked to a summation node dedicated to that particular target class.

During operation, the input pattern, $\boldsymbol{x}$, is first fanned-out to the pattern layer where each pattern unit computes a distance measure between the input and the weight pattern represented by that node. The distance measure (*e.g.* dot-product) is then transformed by a Parzen kernel function. Outputs from the Parzen kernels are summed by the summation nodes. These outputs correspond to estimates of the pdfs of the input pattern with respect to each target class, *i.e.* $P(\boldsymbol{x}|A), P(\boldsymbol{x}|B)$. These probability estimates will be utilised for the combination of predictions from multiple classifiers as presented in section 3.

2.3 Probabilistic Fuzzy ARTMAP

One disadvantage of the PNN is that it encodes every input pattern as a new node in the network, thus increases the network complexity and computational cost if large or unbounded data sets are used. Nevertheless, this problem can be alleviated by using a clustering technique such as FAM.

Our studies have found that there is a close similarity in the network topology between FAM and the PNN. Notice that in Figures 1 and 2, the F_{1a} and F_{2a} layers correspond to the input and pattern layers whereas the map field layer (F_{ab}) corresponds to the summation layer. In one-from-N classification, each node in F_{2a} is permanently associated with only one node in F_{ab}, which is then linked to the target output in F_{2b}. Thus, the map field nodes can be used to sum outputs from all the F_{2a} nodes corresponding to a particular target class, taking the role of the PNN summation nodes.

In view of the suitability of the incremental learning property and the similarity of the network topology between FAM and the PNN, a novel hybrid network, based on the integration of a modified version of FAM [9] and the PNN, has been proposed for on-line classification and probability estimation tasks, and is called Probabilistic Fuzzy ARTMAP (PFAM) [1]. The on-line PFAM algorithm is divided into two phases. First, the FAM clustering procedure is used for classifying the input patterns into different categories (learning phase). Subsequently, the PNN probability estimation procedure is used to predict a target output (prediction phase). The advantage of this integration is two-fold: (i) a probabilistic interpretation of output classes is established which enables the application of Bayes, risk-weighted, classification in FAM; (ii) the number of pattern nodes in the PNN is reduced by the clustering procedure of FAM.

The above description provides a conceptual framework for incorporating FAM and the PNN into a unified, hybrid system, and the rationale behind their integration. In practice, several modifications are necessary to allow effective combination of both the networks, and to increase generalisation ability of the hybrid system. These include methods for estimating kernel centres and widths, and for pre-processing the inputs using fuzzification technique. An explanation of these procedures can be found in [1, 4].

3 SEQUENTIAL EVIDENCE AGGREGATION

In the field of pattern recognition, researchers have shown that combining decisions from multiple classifiers applied to the same data set can improve the performance of individual classifiers [10]. Here, a multiple neural network system is developed where decisions from

multiple PFAM classifiers are combined so that (i) performance of the resulting system can be enhanced; (ii) an alternative approach for handling on-line learning and classification tasks with incomplete data can be realised. Figure 3 depicts a multiple classifier system based on PFAM.

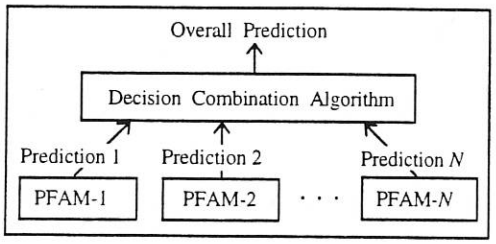

Figure 3 A multiple neural network system

Bayes' rule has been used for reasoning about partial beliefs under uncertainty in multiple classifier systems [10], artificial intelligent systems [11], as well as multiple neural network systems [12]. To understand the Bayesian update of evidence under uncertainty, let H_j, $i=1,\cdots,M$ denote a set of hypotheses, and each H_j be associated with a set of evidences, $e_1,\cdots,e_n$. The posterior probability of the i-th hypothesis can be computed as

$$P(H_i|e_1,\cdots,e_n)=\frac{P(e_1,\cdots,e_n|H_i)P(H_i)}{P(e_1,\cdots,e_n)} \quad (1)$$

where $P(H_i)$ is the prior probability of H_j, and $P(e_1,\cdots,e_n)=\sum_{i=1}^{M}P(e_1,\cdots,e_n|H_i)P(H_i)$ is a normalising factor to ensure the posterior probabilities sum to unity. To enable computation of the posterior probabilities, it is assumed that the pieces of evidence are conditionally independent given H_j (although this assumption may not always be true in every domain). Thus, equation (1) can be simplified to

$$P(H_i|e_1,\cdots,e_n)=\frac{P(H_i)}{P(e_1,\cdots,e_n)}\prod_{i=1}^{n}P(e_j|H_i) \quad (2)$$

One of the attractive properties of Bayes' theorem is its amenability to recursive and incremental computation schemes. The recursive Bayesian update of belief functions is implemented here to aggregate evidence from multiple neural classifiers sequentially. Let the input vector to the j-th classifier be $\mathbf{x}_j$. Upon receiving $\mathbf{x}_j$, classifier j will yield a set of conditional probabilities for M target classes (hypotheses), *i.e.*, $P(\mathbf{x}_j|C_i)$, $i=1,\cdots,M$, which is equivalent to $P(e_j|H_i)$ above. For the first classifier, the posterior probability of class C_i can be computed as:

$$P(C_i|\mathbf{x}_1)=\frac{P(\mathbf{x}_1|C_i)P(C_i)}{\sum_{i=1}^{M}P(\mathbf{x}_1|C_i)P(C_i)} \quad (3)$$

where $P(C_i)$ is the prior probability of C_i. In response to the second piece of evidence from the second classifier, the update of posterior probability can be computed incrementally based on the first evidence as

$$P(C_i|\mathbf{x}_1,\mathbf{x}_2)=\frac{P(\mathbf{x}_2|C_i)P(C_i|\mathbf{x}_1)}{\sum_{i=1}^{M}P(\mathbf{x}_2|C_i)P(C_i|\mathbf{x}_1)} \quad (4)$$

By comparing equations (3) and (4), we can see that the current belief, $P(C_i|\mathbf{x}_1)$ assumes the role of the prior probability in the computation of new belief, $P(C_i|\mathbf{x}_1,\mathbf{x}_2)$. Equation (5) generalises the recursive Bayesian update of belief functions to include the most recent piece of evidence provided by the (j+1)-th classifier,

$$P(C_i|\mathbf{x}_j,\mathbf{x}_{j+1})=\frac{P(\mathbf{x}_{j+1}|C_i)P(C_i|\mathbf{x}_j)}{\sum_{i=1}^{M}P(\mathbf{x}_{j+1}|C_i)P(C_i|\mathbf{x}_j)} \quad (5)$$

Recall that during the prediction phase, the PFAM network uses the Parzen-windows technique to approximate the pdfs. In a multiple classifier platform, the output pdfs from various PFAM classifier thus constitute the supporting evidence for the estimation of the posterior probabilities of target classes. The above Bayesian formalism can then be employed to

combine predictions from multiple classifiers sequentially. During operation, the multiple classifier system utilises whatever information is available to give a set of initial predictions for the target classes. Subsequently, these predictions will be reinforced or counteracted on arrival of new information. Since not all information will be readily accessible for use in many real applications, this multiple classifier system coupled with the recursive Bayesian belief update formalism serve as a simple but effective approach to handle problems with incomplete or missing data.

4 EXPERIMENTS

In the following two sets of experiments, the PFAM network was set to operate at its basic conditions: $\alpha_a \approx 0.0$ (conservative mode); $\bar{\rho}_a = 0.0$ (forced choice); $\beta_a = 1.0$ (fast learning) [2]; overlapping parameter, $r = 2.0$ [1].

4.1 Extension to the "Circle-in-the-Square" Benchmark Problem

The "circle-in-the-square" problem requires a system to identify which points in a square lie inside or outside a circle whose area equals half that of the square. It has been used as a benchmark problem for system performance evaluation in the DARPA artificial neural network technology program [13]. Asfour [14] extends the two-dimensional "circle-in-a-square" task to a three-dimensional case such that the system now needs to identify if a particle is travelling inside or outside a tube residing within a rectangular box (Figure 4). This task is employed to evaluate the effectiveness of FAM and the Fusion ARTMAP network [14] (a modularised ART-based network for multi-sensor data fusion and classification) in handling problems with missing data.

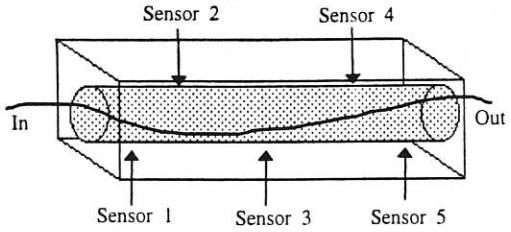

Figure 4 The extended "circle-in-the-square" task requires a system to identify whether a particle is travelling inside or outside a tube residing in a rectangular box.

In the experiment, it is assumed that 5 sensors have been placed along the rectangular box to detect the position of a moving particle as shown in Figure 4. The sensor readings are then fused to 5 different PFAM classifiers to generate a prediction about whether the particle is travelling inside or outside the tube. To simulate a classification domain with missing data, information from the sensors are shut down one by one as if the sensors are faulty.

Algorithm	No. of sensors				
	5	4	3	2	1
Fuzzy ARTMAP	82.24	82.96	83.46	82.28	76.82
Fusion ARTMAP	98.22	97.23	95.86	91.32	88.53
Multiple PFAM System	98.49	98.08	96.49	92.66	86.41

Table 1 Off-line learning results (expressed in percentages) of different number of sensors used in the classification process.

A training set of 100 samples and a test set of 1000 samples were used in the experiment. The experiment was repeated 10 times, and the results were averaged across the 10 runs. Table 1 show the performance of the multiple PFAM system as well as the results of FAM and Fusion ARTMAP as reported in [14]. The results show that the modularised approach of Fusion ARTMAP and the multiple PFAM system outperform the concatenated approach of FAM.

Since the PFAM is capable of learning on-line, a *dual-mode* learning experiment was conducted where each PFAM network was trained, off-line, using 100 data samples. Then, the trained network was engaged in on-line learning using 900 data samples. During the on-line operation, an input pattern was first presented to ART_a with its target output to ART_b. A predicted class was sent from the F_{2a}

winner to ART_b,, and the prediction was compared with the target class to produce a classification result (prediction phase). Then, learning ensued to associate the input pattern with its target class (learning phase). A 100-sample moving window was applied for calculating the on-line accuracy, *e.g.* the accuracy at sample 200 was the percentage of correct predictions from trials 101-200.

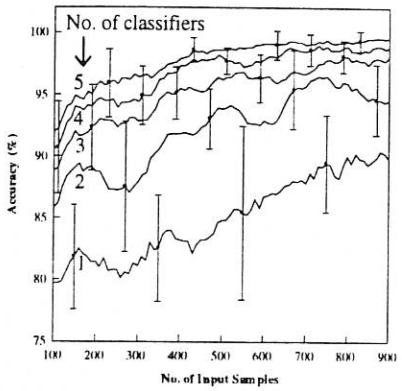

Figure 5 On-line results of different number of classifiers used in the extended "circle-in-the-square" problem

Figure 5 depicts the on-line results averaged over 10 independent runs. The error bars are the standard deviations of the 10 runs to indicate how the results spread across the averaged results. It is clear that the accuracy improves in accordance with the number of classifiers used to give the final prediction, hence justifying the application of multiple neural network systems in classification tasks.

4.2 Diagnosis of Myocardial Infarction

This diagnostic study involved a database of 500 patient records admitted to the Northern General Hospital, Sheffield, United Kingdom, with a major complaint of chest pain (chest pain is known to be strongly associated with Myocardial Infarction (MI) or heart attack). A total of 26 items of electrocardiagraphic and clinical data such as Q waves, ST elevation, age etc. were used as inputs to the multiple PFAM systems. These input features were divided into 4 groups according to their significance to MI (in consultation with a medical expert), and distributed to 4 independent PFAM networks.

In the off-line experiment, the database was divided into a training set of 300 samples and a test set of 200 samples. The most significant feature set formed the inputs to classifier 1, whereas the least significant feature set formed the inputs to classifier 4. Table 2 presents the classification accuracy for the binary decision, MI or not MI, from combining the predictions from multiple PFAM classifiers consecutively. It can be seen that the performance based on the most significant feature set (Classifier 1) was as good as those by combining more than one classifier. In other words, contribution from other input features which are not strongly related to MI would increase the performance slightly, as indicated in Table 2 (the results of 2, 3, and 4 classifiers).

No. of Classifiers			
4	3	2	1
84.1	83.9	82.9	82.8

Table 2 Off-line learning results (expressed in percentages) of the MI diagnosis.

As a comparison, the performance of the admitting clinicians, and the best performances achieved by an MLP network (with optimal decision threshold on a super-set of the same data) [15], as well as by FAM (with voting strategy) [16] are presented in Table 3. The results from the multiple PFAM system are inferior to those of MLP and FAM. Note that the results of MLP and FAM were the best ones after fine-tuning the system parameters. However, the multiple PFAM system was operated at its "basic" settings without any efforts to "optimise" the network parameters.

Algorithm	Accuracy (%)
Clinician	82
MLP	90
FAM	90

Table 3 A comparison with other methods.

A further dual-mode learning experiment was conducted where the multiple PFAM system was trained using the first 100 samples and then tested, along with on-line learning, using the remaining 400 samples. Again, the on-line accuracy was calculated with a 100-sample moving window. Figure 6 indicates the on-line performance, averaged over 10 runs, by combining the predictions from the 4 PFAM classifiers successively. The standard deviations of the 10 runs are plotted as error bars. Unlike Figure 5 where there is a clear improvement in performance with respect to the number of classifiers used, here a single classifier is able to achieve a similar performance comparable to those from more than one classifiers. This phenomenon is understandable as the input feature set to classifier 1 is most significant to the prediction of MI. Nevertheless, as the system encounters more and more samples, it was able to achieve a better performance with a clear delineation between the results by combining various classifiers as demonstrated by the accuracy of the last 100 samples in Figure 6.

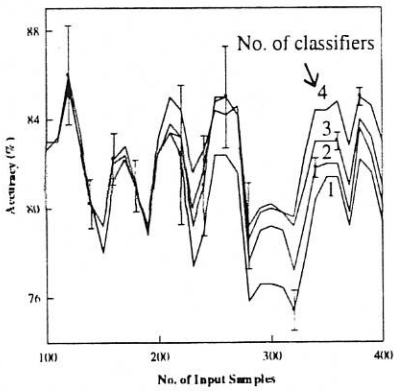

Figure 6 On-line learning results of the MI diagnosis.

5 SUMMARY

A composition of multiple neural network classifiers has been studied to solve pattern classification problems with incomplete or missing data. An autonomously learning, hybrid system of FAM and the PNN network is utilised as the basis for the development of the multiple classifier system. One advantage of the system is the ability to combine decisions from multiple classifier modules sequentially using Bayes' theorem. In this way, the most significant and instantly available data items can be grouped together to a classifier to give an initial prediction. Then, data collected later can be fused to another classifier to reinforce or counteract the initial predictions. As more and more information becomes available, this multiple classifier platform coupled with the sequential decision combination algorithm enables the system to make predictions with greater accuracy as time goes on.

ACKNOWLEDGEMENT

The financial support of the UK EPSRC (Grant No. GR/J43233) is gratefully acknowledged.

REFERENCES

1. Lim, C.P., Harrison, R.F., "An Incremental Adaptive Network for On-line Supervised Learning and Probability Estimation", To appear in Neural Networks.

2. Carpenter, G.A., Grossberg, S., Markuzon, N., Reynolds, J.H., Rosen, D.B., 1992, "Fuzzy ARTMAP: A Neural Network Architecture for Incremental Supervised Learning of Analog Multidimensional Maps", IEEE Trans. on Neural Networks, 3, 698-712.

3. Specht, D.F., 1990, "Probabilistic Neural Networks", Neural Networks, 3, 109-118.

4. Lim, C.P., Harrison, R.F., 1996, "On-line Learning and Probability Estimation with a Self-organising Neural Network", Proc. of the Int. ICSC Symposia on Soft Computing (SOCO'96), B131-B137.

5. Carpenter, G.A., Grossberg, S., 1987, "A Massively Parallel Architecture for a Self-Organizing Neural Pattern Recognition Machine", Computer Vision, Graphics and Image Processing, 37, 54-115.

6. Carpenter, G.A., Grossberg, S., Rosen, D.B., 1991, "Fuzzy ART : Fast Stable Learning and Categorization of Analog

Patterns by an Adaptive Resonance System", Neural Networks, 4, 759-771.

7. Hartigan, J.A., 1975, Clustering Algorithms, New York: John Wiley and Sons.

8. Parzen, E., 1962, "On Estimation of a Probability Density Function and Mode", Annals of Mathematical Statistics, 33, 1065-1076.

9. Lim, C.P., Harrison, R.F., "Modified Fuzzy ARTMAP Approaches Bayes Optimal Classification Rates: An Empirical Demonstration", To appear in Neural Networks.

10. Xu, L., Krzyzak, A., Suen, C.Y., 1992, "Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition", IEEE Trans. on Systems, Man, and Cybernetics, 22, 418-435.

11. Pearl, J., 1988, Probabilistic Reasoning in Intelligent Systems: Network of Plausible Inference, San Mateo, CA: Morgan Kaufmann.

12. Kadirkamanathan, V., Chan, C.P.C., Cheong, K.S. 1995, "Multiple RBF Networks for On-line Identification of Non-liner Systems", Proc. of the fourth IEE Int. Conf. on ANN, 306-311

13. Wilensky, G., 1990, "Analysis of Neural Networks Issues: Scaling, Enhanced Nodal Processing, Comparison with Standard Classification", DARPA Neural Network Program Review.

14. Asfour, R.Y., 1995, "Fusion ARTMAP: Neural Networks for Multi-sensor Fusion and Classification", Ph.D. Thesis, Boston University.

15. Harrison, R.F., Marshall, S.J., Kennedy, R.L., 1991, "A Connectionist Aid to the Early Diagnosis of Acute Myocardial Infarction", Proc. of 3rd European Conf. on Artificial Intelligence in Medicine, 119-128.

16. Harrison, R.F., Lim, C.P., Kennedy, R.L., 1994, "Autonomously Learning Neural Network for Clinical Decision Support", Proc of the Int. Conf. on Neural Networks and Expert Systems in Medicine and Healthcare, 15-22.