NOVEL ART-BASED NEURAL NETWORK MODELS FOR PATTERN CLASSIFICATION, RULE EXTRACTION AND DATA REGRESSION.

YAP KEEM SIAH

UNIVERSITI SAINS MALAYSIA

2010

NOVEL ART-BASED NEURAL NETWORK MODELS FOR PATTERN CLASSIFICATION, RULE EXTRACTION AND DATA REGRESSION.

by

YAP KEEM SIAH

Thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

May 2010

ACKNOWLEDGEMENTS

First of all, I would like to express my thankfulness to my supervisor, Prof. Lim Chee Peng, for his invaluable guidance, motivation, comments, as well as for the fruitful discussions that have resulted in a lot of improvements in this research. Also, I would like to thank him for improving and polishing my writing skills. I am also thankful for his generosity in sharing with me the data sets collected from a power generation plant and a number of hospitals for use in this research.

I would like to express sincere thanks to my co-supervisor, Assoc. Prof. Junita Mohamad Saleh, for her time she spent on review and comments of my journal and conference papers, as well dealing with IPS for urgent matter.

I am greatly grateful to my beloved late parent, Mr. Yap Ah Sa, Madam Khoo Ah Bee, for their endless love, unconditional support and encouragements.

Finally, I would like to thank my friends and colleagues, Dr. Anas Mohammad Quteishat, Dr. Au Mau Teng, Dr. Cheah Cheng Lai, Dr. Chong Kok Hen, Dr. Eric W. M. Lee, Dr. Farrukh Hafiz Nagi, Dr. Izham Zainal Abidin, Dr. Johnny Koh Siaw Paw, Mr. Syed Khaleel Ahmed, Dr. Tan Sing Chiang, Dr. Tay Yong Haur, Dr. Teh Chee Siong and Dr. Tiong Seik Kiong for their sharing of data sets, knowledge and experience.

TABLE OF CONTENTS

CHAPTER 2

INCREMENTAL LEARNING SYSTEMS AND ADAPTIVE RESONANCE THEORY

CHAPTER 3

A GENERALIZED ADAPTIVE RESONANCE THEORY NEURAL NETWORK

CHAPTER 4

AN ENHANCED GENERALIZED ADAPTIVE RESONANCE THEORY NEURAL NETWORK

CHAPTER 5

THE ENHANCED GENERALIZED ADAPTIVE RESONANCE THEORY NEURAL NETWORK WITH RULE EXTRACTION

CHAPTER 6

APPLICATION TO PATTERN CLASSIFICATION PROBLEMS

CHAPTER 7

APPLICATION TO DATA REGRESSION PROBLEMS

CHAPTER 8

CONCLUSIONS AND FURTHER WORKS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACS	Acute Coronary Syndrome - a type of heart attack
ART	Adaptive Resonance Theory
ART-a	Input ART-module in a supervisor ART-based network
ART-b	Output ART-module in a supervisor ART-based network
ARTMAP	Adaptive Resonance Theory with Mapping
BP	Back Propagation
EGART	Enhanced Generalized Adaptive Resonance Theory
EGART-PR	EGART with pruning and rule extraction
EGART-PR-FIS	A Fuzzy Inference System created based on the rules extracted by EGART-PR
EI-ELM-RBF	Enhanced Incremental Extreme Learning Machine with Radial Basis Function
EI-ELM-sigmoid	Enhanced Incremental Extreme Learning Machine with sigmoid additive function
EKF	Extended Kalman Filter
EKFRAN	Extended Kalman Filter type of Resource Allocating Network
ELM	Extreme Learning Machine
eTS	Evolving Takag-Sugeno Model
EUNITE	European Network on Intelligent Technologies for Smart Adaptive Systems
FAM	Fuzzy Adaptive Resonance Theory (ART) with Mapping
FAM-FIS	A Fuzzy Interference System created based on rules extracted by FAM
FAM-RecBFN	Fuzzy ARTMAP - Rectangular Basis Network Function
FAM-RecBFN-FIS	A Fuzzy Interference System created based on rules extracted by FAM-RecBFN
FIS	Fuzzy Inference System
FMM	Fuzzy Min-Max Network
FMM-GA	Hybrid of FMM and Genetic Algorithm
FMM-MACS	FMM-based of Multiple Agent Classifiers Systems

FMM-MACS-FIS	A Fuzzy Interference System created based on rules extracted by FMM-MACS
GAM	Gaussian ARTMAP
GAPRBF	Growing and Pruning Radial Basis Function
GART	Generalized Adaptive Resonance Theory
GGAPRBF	Generalized Growing and Pruning Radial Basis Function
GRNN	General Regression Neural Network
GRNNFA	Hybrid of General Regression Neural Network and Fuzzy ART
$HNFB^{-1}$	Inverted Hierarchical Neuro-Fuzzy Binary Space Partitioning Model
MAPE	Mean Absolute Percentage Error
MFMM1	Version-1 of the Modified Fuzzy Min-Max Network
MFMM1-FIS	A Fuzzy Interference System created based on rules extracted by MFMM1
MFMM1-GA	Hybrid of MFMM1 and Genetic Algorithm
MFMM1-GA-FIS	A Fuzzy Interference System created based on rules extracted by MFMM1-GA
MFMM2	Version-2 of the Modified Fuzzy Min-Max Network
MFMM2-GA	Hybrid of MFMM1 and Genetic Algorithm
MLP	Multi-Layer Perceptron
MRAN	Minimum Resource Allocating Network
O-EGART	Enhanced-Generalized Adaptive Resonance Theory with Ordering algorithm
O-EGART-PR	O-EGART with pruning and rule extraction
O-EGART-PR-FIS	A Fuzzy Inference System created based on the rules extracted by O-EGART-PR
PDF	Probability Density Function
PNN	Probabilistic Neural Network
OSELM	Online Sequential Extreme Learning Machine
OSELM-RBF	OSELM with Radial Basis Function
OSELM-sigmoid	OSELM with sigmoid additive function

OS-Fuzzy-ELM	Online Sequential Fuzzy Extreme Learning Machine
RAN	Resource Allocating Network
RBF	Radial Basis Function
RLA	Real-time Learning Algorithm
RMSE	Root Mean Square Error
SLFN	Single Hidden Feedforward Network
STD	Standard Deviation
SVM	Support Vector Machine
SVR	Support Vector Regression

LIST OF SYMBOLS

Fuzzy ARTMAP

T_j	Choice function of category-j.
$\wedge$	Fuzzy AND operator, $(\mathbf{u} \wedge \mathbf{v})_i \equiv \min(u_i, v_i)$.
$\|.\|$	$\|\mathbf{u}\| \equiv \sum_{i=1}^{M} u_i$
$\mathbf{a}$	M_a-dimensional input vector presented to ART-a.
$\mathbf{b}$	N_a-dimensional class label vector presented to ART-b.
$\mathbf{A}$	Complemented code of $\mathbf{a}$, $\mathbf{A} = (\mathbf{a}, \mathbf{a}^c) \equiv (a_1, a_2, ..., a_M, 1-a_1, 1-a_2, ..., 1-a_M)$.
$\mathbf{x}^a$	Short term memory of ART-a, $\mathbf{x}^a \equiv (x_1^a, ..., x_{2Ma}^a)$.
$\mathbf{y}^a$	Short term memory of ART-a, $\mathbf{y}^a \equiv (y_1^a, ..., y_{Na}^a)$.
$\mathbf{w}_j^a$	The j^{th} ART-a weight vector (long term memory), $\mathbf{w}_j^a \equiv (w_{j1}^a, ..., w_{j,2Ma}^a)$.
$\mathbf{w}_j^{ab}$	The j^{th} weight vector (long term memory) of map field, $\mathbf{w}_j^{ab} \equiv (w_{j1}^{ab}, ..., w_{j,Nb}^{ab})$.
$\mathbf{x}^{ab}$	Short term memory of map field, $\mathbf{x}^{ab} = (x_1^{ab}, ..., x_{Nb}^{ab})$.
ρ_{ab}	User-defined map field vigilance parameter.
$\overline{\rho}_a$	User-predefined baseline vigilance.
ρ_a	The ART-a vigilance parameter.

GART and its enhanced versions

$(\mathbf{A}_k, \mathbf{B}_k)$	$\mathbf{A}_k \in \mathbf{R}^M$ and $\mathbf{B}_k \in \mathbf{R}^L$, are the input vector and kernel label of the k^{th} training sample, respectively.
N	Number of categories
$\boldsymbol{\mu}_j^a$	Center of category-j of ART-a.

$\boldsymbol{\sigma}_j^a$	Standard deviation of category-*j* of ART-a.
n_j^a	Count of category-*j* of ART-a.
$\boldsymbol{\mu}_j^b$	Center of category-*j* of ART-b.
$\boldsymbol{\sigma}_j^b$	Standard deviation of category-*j* of ART-b.
n_j^b	Count of category-*j* of ART-b.
γ_a	Pre-defined initial standard deviation value of ART-a.
γ_b	Pre-defined initial standard deviation value of ART-b.
ε_a	ε parameter of ART-a.
ε_b	ε parameter of ART-b.
$\lambda(.)$	ε-insensitive loss function.
ρ_a	Vigilance parameters of ART-a.
ρ_b	Vigilance parameters of ART-b.
ω	Pre-defined number of cluster centers of ordering algorithm.
$\mathbf{x}$	Unlabeled input vector
$\mathbf{I}_k = (\mathbf{A}_k, 1 - \mathbf{A}_k)$	$\mathbf{I}_k$ is the complemented vector of $\mathbf{A}_k$, where $\mathbf{I}_k \in \mathbf{R}^{2M}$ and $\mathbf{A}_k \in \mathbf{R}^M$.

Post Processing and Rule Extraction

C_j	Confidence factor of category-*j*.
U_j	Usage of category-*j*.
R_j	Accuracy of category-*j*.
$\mathbf{x}_k$	The k^{th} validation sample
$\boldsymbol{\Omega}_k$	Class label of the k^{th} validation sample
q	Number of quantization level.
Q	Total quantization level, Q =5.
V_q	Quantization value.

θ_1	Lower bound of Laplacian likelihood function.
θ_2	Upper bound of Laplacian likelihood function.

FIS

$\wedge$	Fuzzy AND operator, $(\mathbf{u} \wedge \mathbf{v})_i \equiv \min(u_i, v_i)$.
$\vee$	Fuzzy OR operator, $(\mathbf{u} \vee \mathbf{v})_i \equiv \max(u_i, v_i)$.
$m_q(x)$	The q^{th} fuzzy membership function of variable x.
β	Standard deviation of the Gaussian membership function.
N	Number of fuzzy rules.
w_j	Firing strength of rule-j.
$\mathbf{f}_j$	Output constant vector of rule-j.

Others

σ	A common standard deviation used for all pattern neurons of GRNN.
C	Cost parameter of SVM
γ	Factor for the RBF kernel function of SVM.

LIST OF PUBLICATIONS

Journal Papers

Yap, K. S., Lim. C. P. and Abidin, I. Z. (2008) A Hybrid ART-GRNN Online Learning Neural Network with a ε-Insensitive Loss Function. IEEE Transactions on Neural Networks, 19(9), p. 1641-1646, ISSN 1045-9227 (Impact Factor of 2008: 3.726).

Yap, K. S., Lim. C. P. and Mohamed-saleh, J. (2010) An Enhanced Generalized Adaptive Resonance Theory Neural Network and Its Application to Medical Pattern Classification. Journal of Intelligent and Fuzzy Systems, 21(1-2), p. 65-78. ISSN 1064-1246 (Impact Factor of 2008: 0.649).

Yap, K. S., Lim. C. P. and Au, M. T. (2009) A New Generalized Adaptive Resonance Theory Neural Network Model and Its Application to Power Systems. IEEE Transactions on Power Systems. (Under review).

Yap, K. S., Lim. C. P. and Mohamed-saleh, J. (2009) A New Regression Model Based on the Generalized Adaptive Resonance Theory Neural Network and Its Application to Fire Safety Engineering. Expert Systems with Applications. (Under review).

Conference Papers

Yap, K. S., Au, M. T., Lim, C. P, and Mohamed-saleh, J. (2010). Fault Detection and Diagnosis Using An ART-based Neural Network. The 10th International Conference on Artificial Intelligence and Applications, 15-17 Feb. 2010, Innsbruck, Austria.

Yap, K. S., Lim, C. P., Lee, E. W. M. and Mohamed-saleh, J. (2009). Development and Application of An Enhanced ART-Based Neural Network. International Conference on Man Machine Systems ICoMMS 2009, 11-13 Oct. 2009, Pulau Pinang, Malaysia.

Yap, K. S and Lim, C. P. (2008) Short term load forecasting using a modified generalized regression neural network. In Proceeding of IASTED Asia Conference of Power and Energy Systems (AsiaPES), 2-4 Apr. 2008, Langkawi, Malaysia.

Yap, K. S., Lim, C. P., Abidin, I. Z., and Malik, M. (2006) A hybrid neural network for time series prediction. The Third International Conference of Artificial Intelligence in Engineering and Technology, 8-10 Nov. 2006, Kota Kinabalu, Malaysia.

Yap, K. S, Abidin, I. Z. and Lim, C. P. (2006) Short term load forecasting using a hybrid neural network. The First International Conference of Power and Energy, 28-29 Nov. 2006, Putrajaya, Malaysia.

MODEL-MODEL RANGKAIAN NEURAL BARU BERDASARKAN ART UNTUK PENGELASAN CORAK, PENGEKSTRAKAN PERATURAN DAN REGRESI DATA.

ABSTRAK

Tesis ini berkenaan dengan pembangunan model rangkaian neural baru untuk menangani masalah-masalah pengelasan corak, pengekstrakan peraturan dan regresi data. Penyelidikan ini memfokus kepada satu ciri termaju, iaitu kebolehan pembelajaran secara tokokan. Kebolehan ini boleh ditakrifkan sebagai pembelajaran ilmu baru yang beterusan tanpa mengganggu pengkalan pengetahuan yang sedia ada, dan juga tanpa pengulangan set latihan. Model-model rangkaian *Adaptive Resonance Theory (ART)* dan *Generalized Regression Neural Network (GRNN)* bertindak sebagai tulang belakang dalam penyelidikan ini. Hasil model rangkain neural hibrid adalah GART yang berupaya menangani masalah-masalah pengelasan corak, pengekstrakan peraturan dan regresi data. Keupayaan GART juga telah ditingkatkan (dikenali sebagai EGART) dengan beberapa ciri tambahan, iaitu mengunakan fungsi kerugian dan kemungkinan dalam bentuk *Laplacian*, definasi baru fungsi kewaspadaan dan menggunakan mekanisme penjejakan padan . Satu teknik prapemproses pilihan, iaitu algoritma susunan untuk menentukan giliran penyampaian contoh-contoh latihan (dikenali sebagai O-EGART) adalah termasuk. Selepas itu, O-EGART telah ditingkatkan dengan siri pasca pemprosesan (dilambangkan sebagai O-EGART-PR), iaitu, pemangkasan rangkaian dan keupayaan bagi mengekstrakan peraturan dalam bentuk JIKA-MAKA. Prosedur-

prosedur pemangkasan rangkaian memerlukan faktor keyakinan yang dapat dikira berdasarkan satu set pengesahan. Pemberat rangkaian dengan faktor keyakinan yang rendah akan dibuang. Selepas itu, proses pengkuantuman digunakan untuk menukar pemberat yang kekal kepada satu set peraturan dalam bentuk JIKA-MAKA. Sebagai tambahan, satu Sistem Inferensi Kabur (FIS ataupun *Fuzzy Inference System*) telah dibina (dikenali sebagai O-EGART-PR-FIS) untuk tujuan penilai kualiti peraturan-peraturan yang telah diekstrakkan. Prestasi model-model rangkaian neural yang dibangunkan telah dinilai dengan set data bandingan. Kaedah *bootstrap* digunakan sebagai penilai dan pembandingan dengan prestasi pendekatan-pendekatan lain. Bagi tujuan menilai kebolehgunaan praktikal model rangkain neural ini, eksperimen-eksperimen berdasarkan tujuh set data dikumpul dari dunia sebenar, yang gabungan tiga daripada sistem tenaga kuasa, tiga daripada kejuruteraan keselamatan kebakaran dan satu daripada applikasi perubatan, telah dijalankan. Sebagi contoh, kadar ketepatan adalah 98.92% bagi pengelasan arus harmonik dalam rangkaian pengagihan dan 97.20% bagi diagnosis untuk sistem air pengedaran dalam loji penjanaan kuasa, mencadangkan bahawa keupayan model-model rangkaian yang dibangunkan adalah setanding (jika tidak lebih baik) dengan pendekatan-pendekatan lain.

NOVEL ART-BASED NEURAL NETWORK MODELS FOR PATTERN CLASSIFICATION, RULE EXTRACTION, AND DATA REGRESSION.

ABSTRACT

This thesis is concerned with the development of novel neural network models for tackling pattern classification, rule extraction, and data regression problems. The research focuses on one of the advanced features of neural networks, i.e., the incremental learning ability. This ability relates to continuous learning of new knowledge without disturbing the existing knowledge base and without re-iterating through the training samples. The Adaptive Resonance Theory (ART) and Generalized Regression Neural Network (GRNN) models are employed as the backbone in this research. The resulting hybrid neural network model (denoted as GART) is capable of handling pattern classification and data regression problems. The capability of GART is further enhanced (denoted as EGART) with a number of features, which include the used of Laplacian loss and likelihood functions, a new definition of vigilance function, a match tracking mechanism. In addition, a pre-processing technique, i.e., the ordering algorithm, for determining the presentation sequence of training samples is applied (denoted as O-EGART). The O-EGART model is equipped with a series of post-processing procedures (denoted as O-EGART-PR), i.e., network pruning and rule extraction. Network pruning requires computation of the confidence factor of each protoptye node in O-EGART-PR based on a set of validation samples. A quantization process is also applied to convert the prototype weights into a set of IF-THEN rules. In

addition, a standard Fuzzy Inference System (FIS) is constructed (denoted as O-EGART-PR-FIS) in order to evaluate the quality of the extracted rules. The performances of the proposed ART-based models are compared with those from other approaches using benchmark data sets, and the bootstrap method is used to quantify the results. To evaluate the practical applicability of the proposed ART-based models, empirical experiments based on seven benchmark and real-world data sets, i.e., three from power systems, three from fire safety engineering, and one from medical application, are conducted. These results show good performances, e.g., accuracy rates are 98.92% and 97.20% for classification of harmonic currents in distribution network and diagnosis of circulating water systems in power generation plant, respectively, hence justified the usefulness of the proposed ART-based models in undertaking pattern classification and data regression problems.

CHAPTER 1

INTRODUCTION

1.1 Preliminaries

For the last few decades, researches in both theoretical and experiments aspects for the human brain have received much attention. The results indicate that the human brain has a massively parallel architecture composed of many individual simple processing elements (neurons) with intense interconnections (synapses). Generally, early investigations into the human brain were conducted mainly by neurologists, psychologists, and physiologists who developed artificial models for biological nervous systems. However, with the rapid advancements in computing technologies, researches on the artificial brain models, known as artificial neural networks (or simply neural networks), have become popular and have been conducted by researchers from various fields including mathematics, physics, and engineering.

In general, there are two main research interests in neural networks: (i) mathematic modeling of biological nervous systems at the microscopic level of neurons and synapses; and (ii) development of machine learning algorithms that mimic the operation of the human brain at the macroscopic level, whereby the algorithms should be able to perform as intelligently as the human brain in certain aspects such as reasoning, processing information, and inferring decisions.

From the perspective of classification and regression theories, many neural network models can be viewed as extensions of conventional statistical techniques

which have been developed over several decades for undertaking pattern classification and regression problems. The statistical principles embedded in neural networks provide a strong theoretical foundation for the implementation of neural network models as a pattern classifier and/or a data regressor.

Many other characteristics of neural networks have been extensively studied. However, one domain that receives less attention, and yet is important for genuinely *intelligent* learning systems, is the ability to learn new information continually and autonomously without corrupting or forgetting previously learned information. This ability is often referred to as *online learning*, and it is an essential property for a learning system to operate in a non-stationary environment. Note that throughout this thesis, the term *online learning* is used interchangeable with *incremental learning and sequential learning*. The characteristics of this learning strategy are as follows (Huang et al., 2005; Liang et al., 2006; Andonie and Sasu, 2006).

(i) ability to conduct one-pass learning through all data samples, with no re-iteration through the training set.

(ii) ability to learn using only the newly arrived data sample, instead of all past samples, at any time of the training cycle.

(iii) ability to learn new knowledge from the data samples on a one-by-one basis without disturbing the existing knowledge base.

(iv) ability to predict the target output for a new (unlabeled) data sample at any time during the training cycle.

The main aim of this research is to devise an online learning-based neural network model that is able to solve pattern classification and data regression

problems and, at the same time, to extract domain knowledge from the learned network model for explaining its predictions. Two main aspects of neural networks are focused: (i) the online learning property; and (ii) the probabilistic property. Based upon existing neural network models, a number of novel neural network models that integrate the two properties into a common framework are proposed. In order to evaluate the capabilities and applicability of the proposed models, numerous experimental studies using benchmark as well as real-world data sets from various application domains are conducted, with the results compared, analyzed, and discussed.

In the following sections, a definition of and an introduction to neural networks and its applications to pattern recognition and knowledge extraction are provided. A review of the current neural network models for pattern recognition is presented, and motivations for developing the new proposed neural network models are described. Then, the research objectives and scopes are defined, and an overview of the organization of this thesis is included at the end of the chapter.

1.2 Neural Networks

The rapid development of computing technologies have encouraged and inspired advanced researches related to the human brain. The availability of the digital computer as a research tool has tremendously accelerated scientific progresses in many research fields that are very important for understanding the human brain. A conventional computer solves a problem by using an algorithmic method whereby the computer follows a set of instructions. Such an approach requires the computer to know the specific steps to solve a problem. On the other hand, a neural network,

as inspired by biological nervous systems, works with a different paradigm as compared with the conventional computer. The unique element of the neural network is that it comprises a large number of interconnected processing units (known as neurons) working in parallel to solve a specific problem. Indeed, the neural network is a computational method that attempts to simulate (in a gross manner) the biological nervous system of the human brain with two important properties (Graupe, 1997):

(i) It has a self-organizing feature and a learning ability that allow it to solve a wide range of problems.

(ii) It uses simple computational operations to solve a complex, mathematically ill-defined, non-linear, and stochastic problem.

These properties are very similar to the ability of the human brain in solving a problem. A good definition of the neural network is provided by the DARPA (1988) study, as follows.

> "*A neural network is a system composed of many simple processing elements operating in parallel whose function is determined by network structure, connection strengths, and the processing performed at computing elements or nodes.¨Neural network architectures are inspired by the architecture of biological nervous systems, which use many simple processing elements operating in parallel to obtain high computation rates*."

Several computational formalisms of neural network have been developed to handle real-world situations. They are particularly useful in ill-defined and noisy situations. Under these situations, the neural network is more effective and economical as compared with the traditional computational method. For solving

problems arise from non-stationary environments, the learning properties of neural network should be made adaptive. From the point of view of generalization, the neural network has the ability to deal with subsets of the problem domains that are yet to be fully encountered. Otherwise, it is just similar to a mere look-up table that solves a problem based on hard-mapping.

The first neuron model was developed by McCulloch and Pitts (1943 & 1947) and further enhanced by Hebb (1949) with concept of adapting connections between nodes. The Perceptrons model developed by Rosenblatt (1958) was the first artificial neuron model that is capable of performing learning and classification of patterns using simple connections called weights. The Adaline model developed by Widrow and Hoff (1960) has a similar concept as that of Perceptrons, but with the ability to handle data regression tasks. Then, a series of important developments in the area of neural network models has arisen, i.e., the discovery of associative memory (Taylor, 1956), model of self-organization of feature detectors (von der Malsburg, 1973), and ordered neural connections (Willshaw and von der Malsburg, 1976). Later, a number of pioneering studies concerning various properties of different neural network models have been published. These include the Hopfield Network (Hopfield, 1982), the Self Organizing Map (Kohonen, 1982), field theory of self-organizing neural nets (Amari, 1983), back-propagation learning (Rumelhart et al., 1986), and Adaptive Resonance Theory (Carpenter and Grossberg, 1987a, 1987b). All these models have provided a more refined depiction of the brain function than what was anticipated a few decades ago.

Researches in neural network models have found promising results, and these models have been used as a tool for solving problems in various disciplines of science and engineering in the last two decades. In power systems, neural network models have been wide used in many applications, e.g., short-term power load forecasting (Peng et al., 1992; Hippert et al., 2001; Amjady 2007), long-term power load forecasting (Kandil et al., 2002; Carpinteiro et al., 2007), electricity price forecasting (Nogales, 2002; Catalão, 2007), fault diagnosis of power transformer (Zhang et al., 1996; Huang, 2003; Castro and Miranda, 2005), and power system stabilization (He and Malik, 1997; Segal, 2000; Mishra, 2006). In medical applications, neural networks have been used for classification of medical information and diagnosis of diseases (Gletsos et al., 2003; Wei et al., 2005; Lisboa and Taktak, 2006; Serpen, 2008; Erol, 2008; Oğulata et al., 2009). In financial and business, application of neural networks covers credit risk assessment (Jagielska and Jaworski, 1996; Lee and Chen, 2005; Tsai et al., 2009) and forecasting of financial series (Saad, 1988; Koulouriotis et al., 2005; Ghazali et al., 2008). In systems engineering, neural networks have been used for advanced modeling and control, e.g., in aircraft operations (Suzuki et al., 2006; Mori et al., 2007).

1.3 Pattern Recognition

Pattern recognition is an activity that humans perform daily without much conscious efforts. Humans receive patterns (e.g. in visual and audio forms) via sensing organs, whereby the patterns acquired is processed by the brain to form useful information, and subsequently a decision for action to be taken for the patterns is made (Duda et al., 2002). However, this task is not a trivial one for a computerized system. In order to tackle pattern recognition problems, it is necessary for a computerized

system to have techniques and algorithms that are able to process and recognize patterns from data and/or information supplied to the system. Indeed, researches in pattern recognition are conducted by researchers from many disciplines owing to its cross-fertilization nature, which include engineering, computer science, physics, mathematics, and cognitive science.

In general, the task of pattern recognition can be divided into two stages, as shown in Figure 1.1 (Fu, 1968; Tou and Gonzalez, 1974; Young and Calvert, 1974; Duda et al., 2002):

(i) *feature extraction*–finding and extracting a set of significant feature from an input pattern, and then transforming the input features by using some arbitrary function so as to provide informative measurements for the input pattern;

(ii) *classification*–designing a procedure for discriminating the measurements taken from the extracted features, and then assigning it to one of the target classes (classification) or to produce an estimate value (regression) by applying some decision rule.

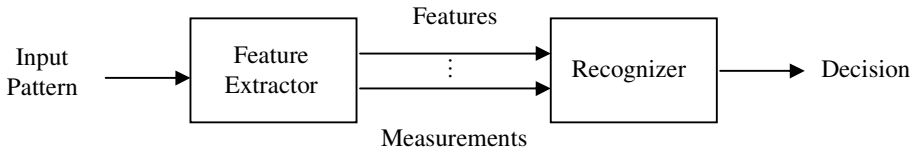

Figure 1.1 A pattern recognition system comprises a feature extractor and a recognizer

This research is focused primarily on the second stage of the recognition process–the classification module. The fundamental problem is to develop a supervised learning procedure which is applicable to a set of data samples (extracted feature measurements) in such a way that each sample is assigned to one of a set of

pre-defined classes (pattern classification) or each sample produces an estimated output value (data regression).

1.4 Rule Extraction

One of the common criticisms of neural network models is that the decision making process is difficult to be understood. In a trained neural network, the knowledge obtained during the training cycle has been parallelly distributed and stored in the network weights. Since the mapping between the input and output spaces learned by the neural network can be nonlinear and non-monotonic (Krishnan et al., 1999), it is obvious that without some form of explanation capability to justify the prediction, the full potential of a trained neural network cannot be realized. Hence, it is useful and important that an explanation capability becomes an integral part of the functionalities of a trained neural network such that its predictions can be explained and justified to the users. This inevitably leads to a higher degree of user acceptance towards the neural network, and to enhance the overall usability of the neural network as a learning and decision making tool. Other than explaining the results, rule extraction of a trained neural network is useful for data exploration and feature revelation, and is able to assist an experienced user to look into a set of conditions under which generalization failure occurs (Qteishat and Lim, 2008).

1.5 Problems and Motivations

The main thrust of researches in classification has been in the use of feedforward neural networks, such as the Multi-Layer Perceptron (MLP) network trained with error back propagation, or other gradient based algorithms (Rumelhart et al., 1986; Gori, and Maggini, 1996; Fine and Mukherjee, 1999; Wu et. al., 2005; Zhang et al.,

2008) and the Radial Basis Function (RBF) network (Broomhead and Lowe, 1988; Moody and Darken, 1989, Plat, 1991; Chng et al., 1996; Schilling et al., 2001; Karayiannis and Randolph-Gips, 2003; Huang et al., 2006a & 2006b; Pedrycz et al., 2008), as pattern recognition and data regression tools. There are a number of attractive properties of such networks for undertaking classification and regression tasks. Cybenko (1989), Barron (1993), Chen and Chao (2009) argue that network architectures using logistic functions are able to approximate any smooth function, under some mild conditions, to an arbitrary degree of accuracy. A similar finding is also concluded for RBF networks where it can approximate any multivariate continuous functions when given a sufficient number of radial basis function units (Poggio and Girosi, 1990; Light, 1992; Huang et al., 2006b & 200c; Huang and Chen, 2008).

From the above arguments, it seems that feedforward networks are useful tools for developing pattern classification and data regression systems. In many aspects, they are. Although theoretical results indicate the capabilities of these networks, there are a number of practical difficulties owing to the network configuration and learning methodology. A problem that often arises is how to determine the optimal number of nodes in the hidden layer(s) (Fujita, 1992; Wu et. al., 2005; Liang et al., 2006). Normally, without any prior information pertaining to the underlying statistics of the data environment, one often has to resort to empirical methods, such as trial-and-error (Wu et. al., 2005; Liang et al., 2006), to obtain a good network configuration for a particular task. Then, some validation test is performed to assess the generalization capability or performance of the trained network. This approach is time-consuming and laborious.

Other than the issue relating to the optimal number of nodes in the hidden layer, the MLP network trained with error back-propagation (Rumelhart et al., 1986) suffers from the problem of local minima (Lippmann, 1987). Generally, the learning rule of error back-propagation uses an optimization process with respect to a cost function. During learning, the network adjusts its weights according to the cumulated errors between the actual and predicted outputs in an attempt to minimize the cost function. The landscape of the error-weight space often consists of a global minimum and some local minima. Thus, it is possible for the learning process to be trapped in a local minimum instead of the desired global minimum. If this happens, the performance and accuracy of the trained network is compromised.

Methods for selecting an optimal or suboptimal network structure for the MLP and RBF networks have been introduced (Baum and Haussler, 1989; Kung and Hu, 1991; Odri et al., 1993; Billings and Zheng, 1995a, 1995b; Liu et al., 2002; Lee and Hou, 2002; Huang et al., 2004 & 2005; Ma and Khorasani, 2005; Peng et al., 2006). In addition, many researchers have proposed techniques to avoid local minima (Baba, 1988; Gori and Tesi, 1992; Kappen and Heskes, 1992; Masters, 1993, 1995; Yiu et al., 2001; Fukuoka et al., 1998; Wang et al., 2004; Behera et al., 2006). Even though the issues pertaining to the optimal network configuration and the global minimum have been solved to a certain degree, the applicability of feedforward networks, as well as many other types of learning systems, is still constrained by their learning methodology.

The learning procedure in most neural networks is essentially an *off-line* process that consists of a training cycle and a test cycle using some data samples.

This approach is useful only when the data environment is stationary, and provided that the training samples are sufficiently representative. This is because during training cycle, information provided by the training samples collected from the environment is encoded by the adjustment (learning) of the network weights. After validating the network performance, the network is put into operation, and no further weight adaptation (or learning) takes place. When the network is presented with an unseen sample, a built-in mechanism for the network to recognize the novelty is not available. In order to learn new information, the network needs to be re-trained using the new sample, together with all previous samples. This is a major drawback in most neural network models, and it arises from the so-called *stability-plasticity* dilemma (Grossberg, 1980; Carpenter and Grossberg 1987a). The dilemma underlies a series of questions, i.e., how a learning system is able to remain plastic or adaptive in response to significant events, and yet remain stable in response to irrelevant ones; how a learning system is able to adapt to new information without corrupting or forgetting previously learned information (Carpenter and Grossberg 1987a, 1988).

This stability-plasticity dilemma has also been termed as the *sequential learning* problem (McCloskey and Cohen, 1989; Ratcliff, 1990). Using the sequential learning approach whereby training is completed for one sample before a new sample is introduced, it is found that a phenomenon known as *catastrophic forgetting* occurs in networks with backpropagation learning. When it happens, newly learned information catastrophically interferes with, and overwrites, previously learned information (McCloskey and Cohen, 1989; Ratcliff, 1990; French, 1991, 1992; Sharkey and Sharkey, 1995). For instance, in an attempt to train a network with backpropagation to perform the arithmetic problem of "add +1",

McCloskey and Cohen (1989) discovered that after training the same network to perform "add +2", it had forgotten how to "add +1". Similar interference problems were also experienced by Ratcliff (1990) in simulations to model how the process of recognition works in humans. For instance, when many items were trained sequentially, only the final item was retained in the memory.

In order to overcome the stability-plasticity dilemma, researchers have proposed new neural network architectures as well as learning algorithms (a review of these neural network models is presented in Chapter 2). Among them, Carpenter, Grossberg, and co-workers have developed a family of neural network architectures called *Adaptive Resonance Theory* (ART) (Carpenter and Grossberg 1987a, 1987b, 1990). There are a variety of ART models for unsupervised as well as supervised learning. Unsupervised ART models include ART1 (Carpenter and Grossberg, 1987a), ART2 (Carpenter and Grossberg, 1987b), ART3 (Carpenter and Grossberg, 1990), Fuzzy ART (Carpenter et al., 1991b), and supervised ART models include ARTMAP (Carpenter et al., 1991a), Fuzzy ARTMAP (Carpenter et al., 1992; Carpenter and Grossberg, 1994). The family of ART models is an example of incremental learning neural networks that self-organize and self-stabilize in response to an arbitrary sequence of data samples in both stationary (time-invariant) and non-stationary (time-varying) environments. Each ART network includes a novelty detector that measures against a threshold of the similarity between the prototype patterns stored in the network and the current input sample. When the match criterion is not satisfied, a new node (neuron) is created with the input sample coded as its prototype. As a result, the number of nodes grows with time, subject to a novelty criterion. Since different tasks demand different capabilities from the

network, this dynamic network architecture and incremental learning methodology avoid the need to have a pre-defined static network size, or to re-train the network with the entire data samples in non-stationary environments.

There are some practical advantages of using incremental (or online, or sequential) learning systems in real-world applications. Many tasks often require system portability and adaptability owing to local differences or the non-stationary nature of the operating environments, such as policy changes, geographical or demographical variations, and advances in new technologies. This means that a static learning system trained on data from a previous site is unlikely to perform optimally using data from a new site due to variations in local conditions. Thus, it is desirable if such a system can be adapted to its changed operating conditions by performing incremental learning of cases from the new site.

From the computational point of view, an incremental learning system offers an extra benefit, i.e., learning can be achieved on the fly in a one-pass process, *i.e.*, each data sample is presented to the network only once. This approach reduces the computational time as the learning system does not need to go through the training samples repeatedly. In addition, the storage demand is eased as the approach does not need to keep all the samples in the memory of the computer.

As compared with an expert system, a neural network is poor in terms of explaining its reasoning process. A definition of rule extraction from a trained neural network as given by Craven and Shavlik (1994) is: “given a trained neural and the examples used to train it, produce a concise and accurate description of the network”.

To a certain extent, the work on devising an autonomous neural network with a rule extraction capability conducted in this research is inspired and motivated by this definition. Indeed, the explanation facility of a neural network is an attractive property for the end user. Therefore, it is essential to equip the neural network with a rule extraction capability in order to provide explanation for its reasoning and predictions. In this context, the supervised ARTMAP network (Carpenter et al., 1991b) has been endowed with such a capability, based on its knowledge representation. The rules extracted from ARTMAP are "soft", i.e., exact matching between input samples and the weights is not necessary; instead, a reasonably close fit suffices.

1.6 Research Scopes and Objectives

The incremental learning methodology of ART (as well as other online learning neural network models) constitutes the backbone of the research in this thesis, and motivates the development of new network architectures and the associated learning algorithms in an attempt to address the stability-plasticity dilemma. In essence, the scope of this research focuses on two areas:

(i) development of neural network-based learning systems that are capable of acquiring knowledge incrementally in both stationary and non-stationary environments with as little supervision as possible;

(ii) development of effective strategies for application of such learning systems coupled with the rule extraction capability to pattern classification and data regression tasks.

A specific supervised ART network, namely Fuzzy ARTMAP (FAM) (Carpenter et al., 1992), is extensively studied in this research. In addition to its *growing* architecture, FAM offers an extra feature as it is a hybrid network combining the advantages of a neural network and fuzzy logic. This integration brings the low-level processing and learning of a neural network and the high-level reasoning of fuzzy logic into a common framework. However, similar to any other systems, FAM is not free from limitations. One phenomenon of the human learning behaviors is that experience gained at the early stage lays a foundation for the knowledge accumulation process in the long run. The same principle applies to incremental learning systems, i.e., the long term performance depends on the sequence or order of training samples. Different sequences of data samples result in different *knowledge bases* in an incremental learning system, hence different performance scores (Carpenter et al., 1992). Further investigation is needed to make the performance of FAM less sensitive towards the order of data presentation.

In addition to FAM, another supervised ART model, namely Gaussian ARTMAP (GAM) (Williamson, 1995), is employed in this research. GAM is a synthesis of an ART network and a Bayesian classifier (Williamson, 1995). The learning algorithm of GAM is similar to FAM, but fuzzy logic equations deployed in FAM are replaced with Gaussian Bayesian equations. Indeed, the learning algorithm of FAM is deterministic in nature. However, pattern classification and data regression problems have been widely studied using statistical theory such as discriminate analysis and Bayesian decision theory (Fu, 1968; Fukunaga, 1972; Duda and Hart, 1973). These statistical approaches offer strong theoretical as well as practical foundations for the implementation of classification and regression systems. The Bayesian classification rates are generally accepted as the optimum results in

terms of quantifying the performance of classifiers in a statistical sense. Assignment of risk factors is also made possible within the Bayesian framework. Therefore, it is worthwhile to investigate how to incorporate the statistical properties into the learning algorithm of FAM.

Apart from investigating the theoretical and algorithmic aspects, effective operational strategies are envisaged for practical application of ART-based model. In summary, this research is geared towards achieving the following objectives:

(i) to develop novel architectures and learning algorithms for incremental learning systems based on ART and Bayesian theorem;

(ii) to investigate the use of an order algorithm to mitigate the effects of sequences of training samples in the developed ART-based models;

(iii) to devise a novel pruning strategy for the developed ART-based models;

(iv) to design a novel rule extraction method from the developed ART-based models;

(v) to demonstrate the applicability of the developed ART-based models to pattern classification and data regression tasks

During the course of achieving the objectives, extensive empirical studies are conducted using benchmarks as well as real-world data sets to evaluate the ART-based models developed in this research. The benchmark data sets are taken from public domain repositories so that performance comparisons with other approaches can be conducted. Besides, a number of real-world data sets collected from industrial organizations are used to demonstrate the applicability of the proposed ART-based models in real environments.

1.7 Thesis Outline

This thesis is organized in accordance with the research objectives. In Chapter 2, a literature review on incremental learning systems by various approaches is presented. Then, incremental learning systems based on ART are introduced. In particular, two variants of the ART networks that are used as building blocks for the new ART-based models developed in this research are examined in detail.

A novel hybrid ART-based model is proposed in Chapter 3 for online pattern classification, probability estimation and regression tasks. A number of simulations based on benchmark pattern classification and data regression tasks are conducted. The results are compared with those obtained by other approaches. The bootstrap method is employed to quantify and compared the results statistically.

In Chapter 4, improvements to the ART-based model developed in Chapter 3 are presented. These include an ordering algorithm (Dagher et. al., 1999) that mitigates the problem associated with sequence of training samples in FAM. The ordering algorithm was originally proposed for tackling classification tasks using FAM. But, in this research, it has been extended to handling data regression tasks.

In Chapter 5, further improvements to the proposed ART-based model (in Chapters 3 and 4) are explained. These include effective post-processing procedures, *i.e.*, confidence factor, pruning, quantization, rule extraction and evaluation of the rules using a classifier based on fuzzy inference systems. Again, extensive experimental studies using benchmark problems are conducted, with the results compared with those obtained by other approaches.

To demonstrate the applicability of the proposed ART-based model (in Chapters 3 to 5) as a pattern classification and data regression tool with a rule extraction capability, benchmark and real-world case studies are presented in Chapters 6 and 7, respectively. Two problems in power systems, one in fire safety engineering, and one in medical diagnosis are considered in Chapter 6. For data regression tasks, three case studies, with one in power systems and two from fire safety engineering, are examined in Chapter 7. The results from all experimental studies in Chapters 6 and 7 are compared with those obtained by other approaches as well.

Finally, conclusions are drawn and contributions of this research are set out in Chapter 8. A number of areas to be pursued as further work are suggested too.

CHAPTER 2

INCREMENTAL LEARNING SYSTEMS AND ADAPTIVE RESONANCE THEORY

2.1 Introduction

The nature of incremental learning is that the learning system keeps updating its knowledge base as a new input sample arrives without having to consider all previous samples. According to Fu (1994), this learning strategy is both biologically and psychologically plausible. Jean Piaget, a noted learning theorist, argues that the external world is built by sequential conceptualization and abstraction of the environment during the early stage of a child's development (Piaget, 1953). Children first grow into their surroundings by direct action; then they draw analogies from concrete examples; later they gradually develop abstract and formal reasoning skills.

From the machine learning point of view, an incremental learning system should be able to differentiate between spurious and rare but important information. Hence, generalisation and selective learning are two main issues. On arrival of a new sample, the system has to decide either to absorb (assimilate) the sample by generalising its knowledge base or to encode (accommodate) the sample into one of the existing information representations (e.g. an existing pattern prototype) in the knowledge base. Indeed, as pointed out by Hrycej (1992), the stability-plasticity dilemma can be viewed as a reformulation of Piaget's theory of assimilation and accommodation in the human developmental stage.

The next section presents a review of incremental learning systems with dynamic structures and learning algorithms. Then, the architecture of Adaptive Resonance Theory (ART) family of neural networks is introduced. In particular, the Fuzzy ARTMAP (FAM) network, which is the backbone of this research, is described in detail. The importance of rule extraction is also explained. A rule extraction technique for ARTMAP-based networks, which was proposed by Carpenter and Tan (1995), is described. A summary is included at the end of this chapter.

2.2 Review of Incremental Learning Systems

The review of related literature covers a number of different approaches for incremental or sequential learning. First, several types of classical and symbolic learning methods are examined. Then, a survey on a variety of neural network-based incremental learning systems is presented. In the survey, different types of neural network models are grouped based on the network architecture and learning algorithm.

2.2.1 Classical and Symbolic Incremental Learning Approaches

The ground work for analysing sequential pattern recognition problems was first proposed by Wald (1947) with the introduction of the Sequential Probability Ratio Test (SPRT). The idea is that by observing an input pattern, or a measurement of an input pattern, the test has to yield a decision either to make a prediction of the output class, or to request for another observation. This method was modified and generalized by Anderson (1960) and Chien and Fu (1966). Other sequential algorithms included the backward procedure using dynamic programming, the non-

parametric sequential ranking procedure and the sequential Bayes test (Fu, 1968; Melsa and Cohn, 1978). Although the SPRT approach can be used for classification tasks (Young and Calvert, 1974), it is more suited for feature extraction problems.

Sebestyen (1962) described a remarkable work on building a representation of data in the input space using Gaussian kernel functions. The approach utilizes the Euclidean distance between the input sample and the cluster centers to decide the output class that the input sample should belong to. Sorenson and Alspach (1971) proposed a recursive Bayesian estimation technique using the Gaussian sums. The technique aims to approximate the probability density function of the state of a noisy dynamic system conditioned on the available measurement data using a convex combination of Gaussian densities. Similar to the Kalman filter (Kalman, 1960), the method is able to perform online approximation of probability density functions using non-orthogonal basis functions based on data samples taken from the system states.

With regard to the method of Sorenson and Alspach (1971), an Adaptive Mixture Model (AMM) for both supervised and unsupervised classification in dynamic environments was introduced by Preibe and Marchette (1991). The method fits a mixture of Gaussian densities based on data samples recursively. It then performs non-parametric estimate of the probability density functions for computing decision regions without explicit assumptions of the underlying functions. The approach also allows the number of Gaussian kernels to grow with the data samples and the target classes to increase over time. Simulation results show that it is able to achieve a close approximation to the estimated function in both stationary and non-

stationary environments. However, the approach relies on a number of underlying assumptions about the operating conditions, especially in the non-stationary environments.

In the symbolic Artificial Intelligence (AI) research, a series of tree-like learning models based on the incremental concept of formation was studied. EPAM (Feigenbaum, 1963) was one of the earliest incremental concept formation systems used for handling classification tasks. This learning model was later refined by Feigenbaum and Simon (1984). The EPAM algorithm builds a discrimination network consisting of nodes and links to represent its acquired knowledge. When the system encounters a sample, it searches through the network until a terminal node is reached. Two learning mechanisms can take place, either familiarization or discrimination. By familiarization, the sample is absorbed into the current terminal node. Otherwise, the node is discriminated or rejected, which, in turn, leads to a new search phase, or to creation of a new link in the network. The EPAM learning mechanism injects two new ideas into the field of symbolic learning machines. First, a discrimination network architecture for concept learning is introduced, and second, both the classification and learning processes are interweaved together, i.e., if the system is not able to classify the current input, it then learns and absorbs the input into its knowledge base.

Inspired by EPAM, many incremental concept formation systems later emerged, e.g., UNIMEM (Lebowitz, 1985), COBWEB (Fisher, 1987), and CLASSIT (Gennari et al., 1989). UNIMEM organizes knowledge into a concept hierarchy of nodes and links through which it sorts new samples. Learning and classification is

also treated as an entity. In addition to growing, UNIMEM performs pruning in order to remove unreliable concept descriptions. However, it lacks a structured method for deciding between various learning operators, and is dependent on user-specified parameters to make decisions (Gennari et al., 1989).

COBWEB (Fisher, 1987) is another symbolic AI model based on the incremental concept. It builds a concept hierarchy with probability information associated with each concept. Unlike UNIMEM, COBWEB does not allow pruning. Instead, a method to split a class into several new classes, or to merge two classes into one is devised. The distinctive point about COBWEB is that the system has a formal foundation in probability theory. Similar to EPAM, COBWEB takes only nominal attributes. A severe limitation of COBWEB is that all samples have to be retained as the terminal nodes in its concept hierarchy. This approach not only makes the system susceptible to noise, but also leads to the possibility of over-fitting the data. In view of the limitations, an unsupervised learning model called CLASSIT (Gennari et al., 1989) was devised. It uses the same control strategy and operators of COBWEB, but differs in the representation of concepts, samples, and the evaluation function. CLASSIT inspects every attribute during the classification process, even when the attribute has no predictive value. Therefore, it is useful for the system to incorporate the idea of selective attention and to focus on certain attributes that contain important information of the target class in its learning process.

2.2.2 Neural Network Approaches

In this section, a review on incremental neural network models is presented. The surveyed models are divided into three main categories, i.e., multi-layer feedforward

networks, basis and kernel function networks, as well as self-organizing and competitive networks.

(a) Multi-layer Feedforward Neural Networks

In an attempt to address the issue of catastrophic forgetting, French (1991, 1992) argued that forgetting is a direct consequence of distributed representation of information in a standard feedforward network trained with back propagation. It is claimed that one way to maintain generalization while reducing catastrophic forgetting is to use a "semi-distributed" representation. An algorithm that allows a multilayer feedforward network to develop a semi-distributed representation was proposed. A factor is used to compute the correlation between the weight vectors encoded by the hidden nodes. As pointed out by French, the approach could result in a loss of information, and affect generalization of the resulting network. However, Park et al. (1991) and Angulo and Torras (1995) showed that adaptive training could be achieved in non-stationary environments without sacrificing the benefits of distributed representation and, at the same time, avoid the catastrophic forgetting problem.

There are a number of algorithms that create nodes automatically in multi-layer feedforward networks. First, a Tiling algorithm for building a network to classify Boolean patterns with guaranteed convergence was proposed by Mezard and Nadal (1989). The number of layers and the number of hidden nodes in each layer are allowed to increase whenever necessary. Nadal (1989) later introduced a network in which hidden nodes are added one by one until the network is able to converge to a solution for the problem at hand. Variants of the Tiling algorithm were

investigated, e.g., the neural tree classifier by Sirat and Nadal (1990), and the parity-machine by Biehl and Opper (1991). On the other hand, an Upstart Algorithm (Frean, 1990) was proposed to build a network for implementing any Boolean mappings. It is claimed that the resulting network is smaller than those produced by the Tiling algorithm. Later, Muselli (1992) combined a sequential learning procedure with the Upstart Algorithm to construct an incremental two-layer perceptron network.

The Cascade-Correlation algorithm (Fahlman and Lebiere, 1990) is another notable learning approach that builds a architecturally-dynamic multilayer feedforward network. The learning procedure starts with a minimal network and incrementally builds a suitable cascaded structure with as many layers as the number of added hidden nodes. Although the network architecture is dynamic, its training assumes an iterative process using the Quickprop algorithm (Fahlman, 1989). Many researchers later investigated and modified the Cascade-Correlation algorithm to suit various application domains (Yang and Honavar, 1991; Smotroff et al., 1991; Sjogaard, 1992; Karunanithi et al., 1992; Hoehfeld and Fahlman, 1992). On the other hand, Lehtokangas (1999, 2000) proposed a technique similar to Cascade-Correction, i.e., constructive backpropagation (CBP). CBP has the same constructive benefits as Cascade-Correction, but with a simpler implementation and the ability to use stochastic optimization routines. Moreover, CBP can be extended to allow addition of multiple new nodes simultaneously, and can be used to perform continuing structure adaptation automatically. This includes both addition and deletion of nodes.