FEATURE SELECTION FOR THE FUZZY ARTMAP NEURAL NETWORK USING A HYBRID GENETIC ALGORITHM AND TABU SEARCH

TANG WENG CHIN

UNIVERSITI SAINS MALAYSIA

2007

FEATURE SELECTION FOR THE FUZZY ARTMAP NEURAL NETWORK USING A HYBRID GENETIC ALGORITHM AND TABU SEARCH

by

TANG WENG CHIN

Thesis submitted in fulfilment of the requirements for the degree of Master of Science

July 2007

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisor, Associate Professor Dr Mandava Rajeswari and co-supervisor, Associate Professor Dr. Lim Chee Peng for all the insightful guidance and encouragement throughout this work. I am greatly thankful for their invaluable guidance, support, comments, and for many of the fruitful discussions, for which all of these merits lead to many improvements to this research work.

I would also like to thank my parent, Tang Boon Seng and Chong Saw Lan who provide continuous support and encouragement. My special thanks go to Chan Sok Feng for her caring, patience and greatest company.

Last but not least, I wish to thank all my co-worker in USM, whom have offered pleasant, helpful assistance, and whom have lent their expertise, in both aspects of technical knowledge and experiences, to assist me in this research work. I would also thank Dr. Janaka Low for his motivation and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANNs	Artificial Neural Networks
ART	Adaptive Resonance Theory
BP	Backpropagation
CPN	Counter-Propagation Network
DA	Discriminant Analysis
EDFE	Interclass Euclidean Distance
ERFE	Exception Ratio Based Feature Elimination
FAM	Fuzzy ARTMAP
FCS	Fuzzy Classification System
GA	Genetic algorithm
kNN	k-Nearest Neighbour
MDLP	Minimum description length principle
MI	Myocardial Infraction
MLP	Multi-Layer Perceptron
MOGA	Multiple objective genetic algorithm
NN	Neural Networks
NNC	Nearest-Neighbour Classifier
PCA	Principle Component Analysis
PDP	Parallel distributed processing
PID	Pima Indian Diabetes
PNN	Probabilistic Neural Network
RBF	Radial Basis Function
SA	Simulated annealing

PEMILIHAN CIRI BAGI RANGKAIAN NEURAL ARTMAP KABUR DENGAN MENGGUNAKAN SATU HIBRID ALGORITMA GENETIK DAN PENCARIAN TABU

ABSTRAK

Prestasi pengelas rangkaian neural amat bergantung kepada set data yang digunakan dalam process pembelajaran. Secara praktik, set data berkemungkinan mengandungi maklumat yang tidak diperlukan. Dengan itu, pencarian ciri merupakan suatu langkah yang penting dalam pembinaan suatu pengelas berdasarkan rangkaian neural yang efektif. Tesis ini mempersembahkan penyelidikan tentang satu algoritma hibrid dalam pemilihan ciri bagi pengelas rangkaian neural ARTMAP Kabur dengan menggunakan Algoritma Genetik (GA) dan Pencarian Tabu (TS). Algoritma hibrid yang dicadangkan, GA-TS, menggabungkan struktur ingatan baru-baru (recency) dan ingatan kekerapan (frequency) ke dalam proces pencarian GA. Struktur ingatan baru-baru TS membantu meluaskan pencarian GA. Manakala, struktur ingatan kekerapan TS memberi panduan kepada operator genetik dan membantu dalam penyempitan process pencarian GA. Satu siri kajian empirikal yang melibatkan masalah piawai dan masalah dunia sebenar digunakan untuk menilai keberkesanan algoritma hibrid yang dicadangkan. Rangkaian neural ARTMAP Kabur digunakan sebagai pengelas asas dalam kerja penyelidikan ini. Satu kaedah simulasi penyuntikkan ciri berlebihan juga dibangunkan bagi menilai keupayaan GA-TS dalam mengenalpasti dan mengeluarkan ciri berlebihan yang boleh mengurangkan kejituan pengelasan. Keputusan eksperimen menunjukkan sistem GA-TS mempunyai prestasi yang lebih bagi berbanding dengan GA biasa dari segi kepadatan ciri dan kejituan pengelasan.

FEATURE SELECTION FOR THE FUZZY ARTMAP NEURAL NETWORK USING A HYBRID GENETIC ALGORITHM AND TABU SEARCH

ABSTRACT

The performance of Neural-Network (NN)-based classifiers is strongly dependent on the data set used for learning. In practice, a data set may contain noisy or redundant data items. Thus, feature selection is an important step in building an effective and efficient NN-based classifier. In this thesis, the research of a hybrid algorithm of Genetic Algorithm (GA) and Tabu Search (TS) for feature selection in the Fuzzy ARTMAP NN classifier is presented. The proposed GA-TS algorithm embeds the recency and frequency memory structures of TS into the search process of the GA. The recency memory structure helps induce an additional diversification mechanism in the GA search process. On the other hand, the frequency memory structure provides guidance to genetic operator and helps intensify the GA search process. A series of empirical studies comprising benchmark and real-world problems is employed to evaluate the effectiveness of the proposed hybrid GA-TS algorithm. A simulated noisy feature injection method is devised to assess the capabilities of GA-TS in identifying and removing noisy features that can degrade classification accuracy. Experimental results demonstrate that proposed GA-TS performs better in terms of feature compactness (the number of features reduced) and classification accuracy than the ordinary GA.

CHAPTER 1

INTRODUCTION

1.1 Preliminaries

One of the nice features of the human brain is its ability to learn many new things without forgetting things learned earlier. Researches in both theoretical and experimental aspects of the brain have revealed that the human brain is composed of many individual processing elements, known as neurons (Etheridge and Brooks, 1994). The complex, nonlinear, and parallel information processing architecture of these biological neurons play an important role in processing information. Linked with dense interconnections, these neurons have impressive capabilities in performing certain tasks, such as pattern recognition and perception (Bishop, 1995). In the early days, brain researchers were mainly neurologists, psychologists, and physiologists who developed artificial models for biological nervous systems. However, over the past few decades, these artificial models, commonly known as *Artificial Neural Networks* (ANNs), or simply *Neural Networks* (NNs), have become an active area of investigation. To date, this area of research is highly interdisciplinary, and is extensively researched by professionals from various fields including computer science, mathematics, physics, and engineering. Because of the intelligent behaviour of the human brain that can learn many new things, it would be highly desirable if we could impart the same capability to the NN models.

In the following sections, an introduction to NNs and pattern classification is provided. A review on various feature selection methods is presented. Then, the problems and motivation, research objectives, research methodology and scope are

explained, and an overview of the organisation of this thesis is included at the end of the chapter.

1.2 Neural Networks

NNs are relatively crude models of the neural structure of the human brain. According to Marks II (1993), a NN represents a *computational* approach to intelligence, as contrasted with the traditional, more symbolic approaches. The first major contribution on NNs was made by McCulloch and Pitts (1943; 1947). Nevertheless, it was the work by Hebb (1949) that first triggered the concept of adapting the connections between nodes or processing elements, hence learning in NNs. Ever since publication of Hebb's law, a variety of different network architectures and learning paradigms have been proposed. Amongst the earliest models are the "Perceptrons" (Rosenblatt, 1958), the "Adaline" (Widrow and Hoff, 1960), and the Hopfield networks (Hopfield, 1982; 1984). These artificial neural models may or may not be biologically plausible, but they always include connections and nodes analogous to biological nerve nets.

The *DARPA* (1988) study provides a reasonable definition for the term NNs, as quoted:

> "A NN 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. … NN architectures are inspired by the architecture of biological nervous systems, which use many simple processing elements operating in parallel to obtain high computation rates."

Research in NNs has found promising results in many fields, and they have been used as a problem-solving tool in various disciplines of science and engineering.

In the following, a list which represents only a sampling of areas in which NNs have been successfully implemented is presented:

i) Speech recognition (Lin *et al*., 2000)

ii) Handwritten character recognition (Wu *et al*., 2000; Hanmandlu *et al*., 1999)

iii) Personal Identification (Nagaty, 2003; Han *et al*., 2003; Ma *et al*., 2003)

iv) Electrical signal recognition (Engin, 2004; Khandetsky and Antonyuk, 2002)

v) Automatic vehicle control (Ohno *et al*., 1994; El Hajjaji and Bentalba, 2003)

vi) Medical diagnosis (Hayashi and Setiono, 2002; Zhou and Jiang, 2003; Leung and Mao, 2003; Ergun *et al*., 2004)

vii) Detection of explosives (Nunesa *et al*., 2002)

viii) Prediction of bank failure (Tung *et al*., 2004)

ix) Stock market prediction (Kim and Lee, 2004)

x) Prediction of protein secondary structures (Hu *et al*., 2004)

xi) Chemistry (Kewley , 2000; Winkler, 2004)

1.3 Pattern Classification

The task of recognition and classification is one of the most frequently encountered decision making problems in daily activities. A classification problem occurs when an object needs to be assigned into a predefined group or class based on a number of observed attributes, or features, related to that object. Humans constantly receive information in the form of *patterns* of interrelated facts, and have to make decisions based on them. When confronted with a pattern recognition problem, stored knowledge and past experience can be used to assist in making the correct decision. Indeed, many problems in various domains such as financial, industrial, technological, and medical sectors, can be cast as classification problems. Examples include

bankruptcy prediction, credit scoring, machine fault detection, medical diagnosis, quality control, handwritten character recognition, speech recognition etc.

Pattern recognition and classification has been studied extensively in the literature. Among some of the classic textbooks in pattern recognition and classification include Fu (1968), Fukunaga (1972), as well as Duda and Hart (1973). In general, the problem of pattern recognition can be posed as a two-stage process, as shown in Figure 1.1 (Fu, 1968; Duda and Hart, 1973; Tou and Gonzalez, 1974; Young and Calvert, 1974):

(i) ***feature extraction*** – which involves selecting the significant features from an input pattern, and transforming them through some function that can provide some informative measurements for the input pattern;

(ii) ***classification*** – which involves devising a procedure for discriminating the measurements taken from the extracted features, and assigning the input pattern into one of the possible target classes according to some decision rule.

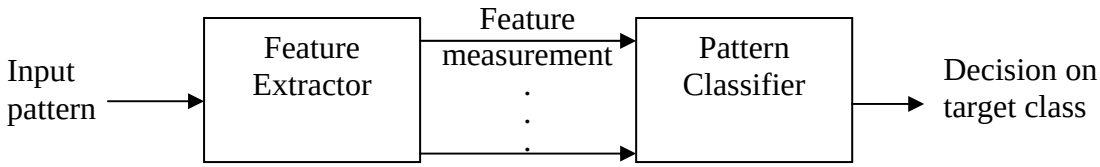

Figure 1.1 A pattern recognition system consists of a feature extractor and a pattern classifier. An input pattern is transformed into a set of measurements by the feature extractor, and assigned to one of the target classes by the classifier using some decision rules.

The research work detailed in this thesis is focused on *feature selection*. The fundamental problem addressed is to devise an automated feature selection algorithm to build a compact and concise data set for pattern classification using NN models.

1.4 Classification Systems

Generally, classification systems can be categorized into several categories. Among the popular pattern classification systems include statistical-based and NN-based classification approaches

1.4.1 Statistical classification

Statistical classification approaches are generally based on probability models. This type of approach is normally used by statistician, whereby the involvement of a statistician is needed in the overall process of structuring the problem. (Mitchie *et al*, 1994).

One of the earliest statistical classification methods is the discriminant analysis. The idea of this method is to divide the sample space using a series of lines to separate each class. Fisher (1936) introduced the first version of discriminant analysis, which was known as Linear Discriminant. Since then, various methods have been developed, e.g. Quadratic Discriminant and Logistic Discriminant.

Other statistical methods include density estimation (Fix and Hodges, 1951) and *k*-nearest neighbour (*k*NN). Kernel density estimation is used in the classification of sample data in density estimation approach. In *k*NN, it finds in the *N*-dimensional feature space the closest object from the training set to an object being classified.

As mentioned earlier, the use of statistical classification methods requires statistical knowledge to define the structure of the problem. In other words, human intervention is required, and different problems may require different settings. This forms the primary weakness of statistical-based methods in pattern classification,

whereby it is not robust, especially for non-statistical users (Mitchie *et al*, 1994). One alternative for robust classification would be NN-based approaches.

1.4.2 Neural network-based classification

Perhaps feedforward networks especially the Multi-Layer Perceptron (MLP) (Rumelhart *et al*, 1986) and Radial Basis Function (RBF) (Broomhead and Lowe, 1988; Moody and Darken, 1989) networks are the most well-known NN-based classifiers. The success of these NN models is based on its capabilities as a universal approximator. Cybenko (1989) argued that network architectures using logistic functions are able to approximate any smooth function, under some mild conditions, to an arbitrary degree of accuracy. Furthermore, a similar finding is also concluded for RBF networks. Poggio and Girosi (1990) and Light (1992) showed that a RBF network can approximate any multivariate continuous functions when given a sufficient number of radial basis function units. Another advantage of feedforward NNs is their strength as a Bayesian probability estimator. The MLP and RBF networks have been shown to exhibit this characteristic, in which their outputs can be regarded as estimates of posterior probability distribution (White, 1989; Wan, 1990; Richard and Lippmann, 1991).

However, although the theoretical results indicate the capabilities of feedforward networks, there are a number of difficulties in practical applicability of these networks owing to the network configuration and learning methodology. For example, a problem that often arises is determination of the optimal number of nodes in the hidden layer(s) (Fujita, 1992). Another problem with the MLP network trained with error back-propagation (Rumelhart *et al*, 1986) is the existence of local minima (Lippmann, 1987). Even assuming that the optimal network configuration and the global minimum are attained, the applicability of feedforward networks, as well as many other networks, is constrained by their learning methodology.

1.4.2.1 The Stability-Plasticity Dilemma

Learning in most NNs is essentially an *offline* process, which consists of a training phase and a test phase, using some data samples, *i.e*. pattern samples. In order to accommodate new information, a previously trained NN has to be re-trained using the newly available pattern samples. This is known as the *sequential learning problem* (McCloskey and Cohen, 1989; Ratcliff, 1990) or the *stability-plasticity dilemma* (Grossberg, 1980; Carpenter and Grossberg 1987a).

In sequential learning, training is done on a sample-by-sample basis, and not on a batch mode. This approach leads to a phenomenon called catastrophic forgetting in back-propagation learning of feedforward networks, i.e., previously learned information is catastrophically overwritten by newly acquired information (McCloskey and Cohen, 1989; Ratcliff, 1990; French, 1991, 1992; Sharkey and Sharkey, 1995).

The sequential learning problem is also addressed as the stability-plasticity dilemma by Grossberg (Grossberg, 1980; Carpenter and Grossberg 1987a). This dilemma poses the fundamental questions in autonomous learning systems, i.e., how a learning system can preserve existing knowledge while continuing learning new information; how a learning system can prevent newly learned knowledge corrupted memories of prior learning. In response to the stability-plasticity dilemma or the sequential learning problem, many researchers have proposed new network architectures and learning algorithms. 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).

The family of ART networks provides several significant advantages over other types of NNs. Among the important features that subscribed by ART to be a successful autonomously learning system include (Downs *et al.*, 1995)

- An ability to discriminate novelty from noise, and familiar events from rare but important ones;
- Fast learning based on predictive success rather on predictive failure (mismatch);
- Self-organisation, with few arbitrary parameter to tune, and automatic structure determination;
- Linear rather than exponential scaling with problem size;
- Straightforward revelation of embedded rule set;
- Inherently parallel implementation.

These features are essential for computational demanding tasks, e.g. feature selection in NN-based classification.

1.4.2.2 Fuzzy ARTMAP

Among various types of ART networks, Fuzzy ARTMAP (FAM) (Carpenter *et al*, 1992) has emerged as a powerful supervised ART-based model for tackling pattern classification problems (Obaidat and Saudon, 1997; Lee and Tsai, 1998; Heinke and Hamker, 1998; Aggarwal *et al*, 1999). FAM combines the salient properties of ART with fuzzy set theory. FAM is very fast in training (Carpenter and Grossberg, 1994; Carpenter *et al*, 1995). As compared with a large number of training epochs needed in other NN model (e.g. MLP), FAM requires relatively few training epochs, which can be conducted incrementally. FAM also is proven to be noise tolerant (Charalampidis and Kasparis, 2001). In the context of pattern classification, FAM has been shown to produce good performance in a number of benchmark classification tasks (Carpenter and Grossberg, 1994,1995). FAM has also been applied to tackle various real-world applications involving pattern classification with good performance, such as medical

diagnostic (Ham and Han, 1996; Azuaje, 2001; Vigdor and Lerner, 2006)), fault detection (Aggarwal *et al*, 1999; De and Chatterjee, 2004; Tan and Lim, 2004), manufacturing decision support system (Tan et al, 2005) and biometrics (Obaidat and Saudon, 1997; Lim and Woo, 2006).

The above characteristics make FAM an attractive NN model for investigation into the problem of feature selection in NN-based classification. Therefore, FAM is selected as the base NN classifier in this research.

1.5 Problems and Motivations

1.5.1 Problems in NN-based Classification

As highlighted earlier, the focus of this research is on the development of an automated feature selection algorithm to build a compact and concise data set for pattern classification using the FAM model. Feature selection is a process of selecting a subset of *n* features from a set of *N* features based on some optimization criterion (Lee *et al*, 2004). In NNs, feature selection has two major functions, i.e. to reduce the complexity of the NN model and to identify important features from a data set. The former objective helps build a concise classification system whereas the latter objective helps eliminate redundant features or noise in the data set used. The ultimate aim is to produce a NN classifier with a good performance in terms of accuracy (Chaika and Yulu, 1999).

Rising of interest in feature selection recently is owing to several reasons. One of the primary reasons is the wide implementation of NN-based classification system in various applications, particularly in pattern classification. In classification applications, the priority is stressed on classification accuracy. The learning mechanism in NN is a form of inductive learning; i.e., learning from specific data to form general rule.

Therefore, the performance of NN is strongly determined by data. During data collection, a user deploys his/her experience to select data attributes or features that are assumed useful for the classification task. In reality, the data set may contain noisy or redundant data without being realized by the user. This may affect classification accuracy. Therefore, there is a need for an automated feature selection algorithm for NN-based classifiers.

The rapid developments of new application of NN models in dealing with vast amount of data such as medical data processing (Puuronen *et al*, 2000), data mining (Piramuthu 1998, Martin-Bautista and Vila 1999), and multi-media information retrieval (Lew 2001, Liu and Dellaert 1998, Messer and Kittler 1997) also contributes to researches in feature selection. Fast processing of large volume of data is critical for these applications that require real-time response. Thus, selecting the important features and limiting the number of feature into a manageable size is an essential requirement.

1.5.2 Motivations in Using Hybrid GA-based Search Techniques

In many NP-hard problems such as feature selection (Hyafil and Rivest, 1976; Blum and Rivest, 1992), the search space is complex and irregular. Traditional search methods e.g. *blind search*, *heuristic search*; often perform less than desirables as they are not robust enough to escape from local minima (Miller *et al* 1993). Hence a robust global search method is needed. In Chaturvedi and Carroll (1997), Pudil *et al* (1994), Law *et al* (2004), Yang and Honavar (1998), various search methods including sequential forward/backward searches, floating search, beam search, bidirectional search, Particle Swarm Optimisation (PSO) and genetic algorithm (GA) are applied to feature selection. To handle irregular and complex search spaces, the search should adopt a global strategy and rely heavily on intelligent randomization. PSO, Ant Colony (AC) and GAs follow just such a strategy. A review on these global search methods is

conducted on the next chapter. Based on the review in section 2.4.4., PSO focuses more on local neighbour information during search, thus inheriting similar weaknesses of local search (Firpi and Goodman, 2004). On the other hand, the review also reveals that AC is inclined more to solving routing, path finding, and decision tree like problems (Dormigo and Gambardella, 1997; Sim and Sun, 2003; Chiang et al, 2006), and is not a common approach to feature selection tasks. Hence, this research is focused on the use of the GA for feature selection.

The appropriateness of using the GA in feature selection problems can be seen from its good performance and wide implementation as documented in the literature (Brill *et al*, 1992; Raymer *et al*, 2000; Jack and Nandi, 2000; Zio *et al*, 2006). However, extensive experimentation and experience from a large number of applications revealed some limitations and shortcomings of GAs. GAs may be efficient in locating the optima in the search space. But, GAs can suffer from excessively slow convergence before finding an accurate solution because of the characteristics of the use of minimal *a priori* knowledge and failure of exploiting local information (Renders and Flasse 1996). To resolve this weakness, GAs have been combined with other search algorithms (Ackley 1987, Goldberg 1989, Kazarlis *et al* 1996, Miller *et al* 1993, Mitchell *et al* 1994, Mühlenbein 1992, Papadakis and Theocharis 1996, Petridis *et al* 1998, Renders and Bersini 1994). Hence, this research focuses on the investigation of a hybrid GA model for feature selection.

The earliest innovation in hybrid GA approaches is the integration of local search method with GA (Miller *et al*, 1993; Dengiz *et al*, 1997; Menozzi *et al*, 1996). In most of the hybrid systems, local search is added into GA as an additional search operator. The intention is to utilize the advantages of neighborhood search in local search methods to improve the GA efficiency. However, adding extra operator implies the increase of computational load in GA. Hence, the current trend has moved towards

hybridizing GA with other *metaheuristic* search methods such as *Simulated Annealing* (SA) and *Tabu Search* (TS). In this approach of hybridization, the concept of each technique is mixed together to complement each other's weaknesses. This leads to the formation of enhanced search strategies without adding extra search operator.

SA can be viewed as an enhanced version of *hill-climbing* search, whereby the main advantage is the usage of the *temperature* parameter in SA to control the degree randomization (Kirkpatrick *et al*., 1983). TS, on the other hand, utilizes various kinds of memory (recency, frequency, quality, and influence) as guidance in the search process (Glover, 1986). In GA, the degree of randomization is controlled via a selection operator. Hence, hybrid GA-SA algorithm caught less attention as compared with hybrid GA-TS systems.

For TS, context forms the fundamental of attributes definition and the determination of move neighbourhoods, and in the choice of conditions to define tabu restriction. GAs, on the other hand, stresses the freedom of its rules from context. Crossover in GAs is a *context neutral* operation, whereby it is assumed to be independence from any condition of that solution must obey in a particular problem setting. In practice, however, it is generally taken this as an inconvenient assumption, which makes the solution of interest difficult to find. Consequently, a good deal of effort is to implement TS's context in GA operations, particularly in crossover and mutation. Such implementation allows genetic operators to remove deficiencies of standard operators upon being confronted by changing context, hence addressing context directly and making it an essential part of the design for generating combinations. Hence this research focused on the integration of TS into GA.

1.6 Research Objectives

In feature selection, a learning system is required as the fitness evaluator. In order to have an efficient search, the learning system must be accurate, fast, and easy to configure. As mentioned earlier, most NN models suffer from these problems, particularly on learning speed and configuration complexity. The FAM network meets the requirements of such a fitness evaluator. Its characteristics of fast learning, noise tolerance, and flexible configuration makes it a suitable fitness evaluator.

All the above reasons drive this research to focus on the investigation of a hybrid algorithm of GA and TS for feature selection, and the incorporation of the resulting feature selection approach in FAM to tackle pattern classification problems. Specifically, the main objectives of this research are as follows:

i. To examine the feasibility of the GA for feature selection in FAM;
ii. To devise a hybrid GA-TS algorithm for feature selection for FAM;
iii. To demonstrate the effectiveness and applicability of the hybrid GA-TS algorithm coupled with FAM in tackling pattern classification tasks.

1.7 Research Scope

As mentioned earlier, FAM is selected as the baseline NN classifier in this research because of its desirable characteristics in tackling the stability-plasticity dilemma and its good performances in handling pattern classification problems. Since feature selection is a NP-hard problem, global search methods are much suitable as compared with local search methods. Hence this research is focused on global search strategies for feature selection. This research is also focused in wrapper-based feature selection techniques owing to it's superiority in terms of classification accuracy. Among various global search strategies, the GA is selected as the search method in this research owing to its strength in formulating a good platform for a parallel search using global information as well as its suitability in handling feature selection problems, as revealed

from the literature review (in section 2.5). The enhancement to the GA search strategy is then introduced by integrating local metaheuristic search methods into the GA, and TS in particular is selected for investigation. From the literature review, combination of GA and TS has been proposed and has shown good results. However, the application of hybridisation of GA and TS to feature selection problems is still new. Therefore, it is worthwhile to research into the effects of integrating TS into the GA with the most basic configuration, and compare the performances empirically on classification accuracy and network compactness. In short, the main scope of this research is to develop an automated feature selection algorithm, based on a hybridisation of GA-TS, for FAM.

To evaluate the capabilities and applicability of the proposed algorithm, a series of empirical studies are performed using benchmark data obtained from public domain repositories. The data sets used include Horse Colic, Thyroid, Iris, Pima Indian Diabetes (PID), Wisconsin Prognostic Breast Cancer (WPBC), Heart disease, Ionosphere, German credit ranking, Single Proton Emission Computed Tomography (SPECTF), and Hepatitis. The selection of data sets used in each experiment is based upon the availabilities of published results for performance comparison purposes. In addition, two real medical data sets, i.e. Myocardial Infraction and the acute stroke diagnoses, are tested to demonstrate the applicability of FAM coupled with the proposed GA-TS algorithm to medical pattern classification tasks.

1.8 Research Methodology

A step-by-step methodology is applied to achieve the research objectives. First, a standard GA is developed using the most basic configuration. The GA parameters, particularly the crossover and mutation rates are set based on the rules of thumb e.g. a higher rate for crossover and a lower rate for mutation. The capabilities and applicability of GA in feature selection are then evaluated using a series of experiment with benchmark data sets i.e. Horse Colic, Pima Indian Diabetes (PID), Iris and Thyroid

data sets, and the results are then compared with other published results. First, an experimental study is conducted to justify the suitability of using FAM coupled with the GA in this research. Then, an experimental study is conducted to demonstrate the effectiveness of using the GA as the search method for feature selection.

The research continues with the proposal of a hybrid GA-TS algorithm for feature selection. Two memory structures of TS are integrated into the GA. The performances of the proposed system are assessed and compared with those from the ordinary GA using WPBC, Heart disease, Ionosphere, German credit ranking, SPECTF, and Hepatitis data sets. A performance comparison with other published results is conducted. The bootstrapping method is also applied to quantify the performance indicators, i.e., classification accuracy and number of noisy features reduced, statistically. The proposed approach is then evaluated using real-world data sets in the medical domain, i.e. Myocardial Infarction (MI) and acute stroke diagnoses. Besides that, studies on the evolution process and the effect of GA-TS in terms of computational demand are conducted.

Next, the research is focused on the evaluation of the noisy feature reduction capability of the proposed GA-TS algorithm. Two methods are devised to inject noise into existing data sets. They are noise injection to existing feature and extra noisy feature injection. The proposed GA-TS algorithm and standard GA are then put to test using WPBC, Heart disease, Ionosphere, German credit ranking, SPECTF, and Hepatitis data sets. The performance comparison is focused on the capabilities of each method in identifying and removing simulated noisy features.

In the last step, real applicability of FAM coupled with the proposed GA-TS algorithm to medical data (MI and stroke) classification is studied. Two additional performance measures, i.e. specificity and sensitivity that are commonly used in

medical data analysis, are employed to quantify the results. Implications of FAM with GA-TS in medical data classification tasks are analysed and discussed.

The overall research methodology is summarised in Figure 1.2

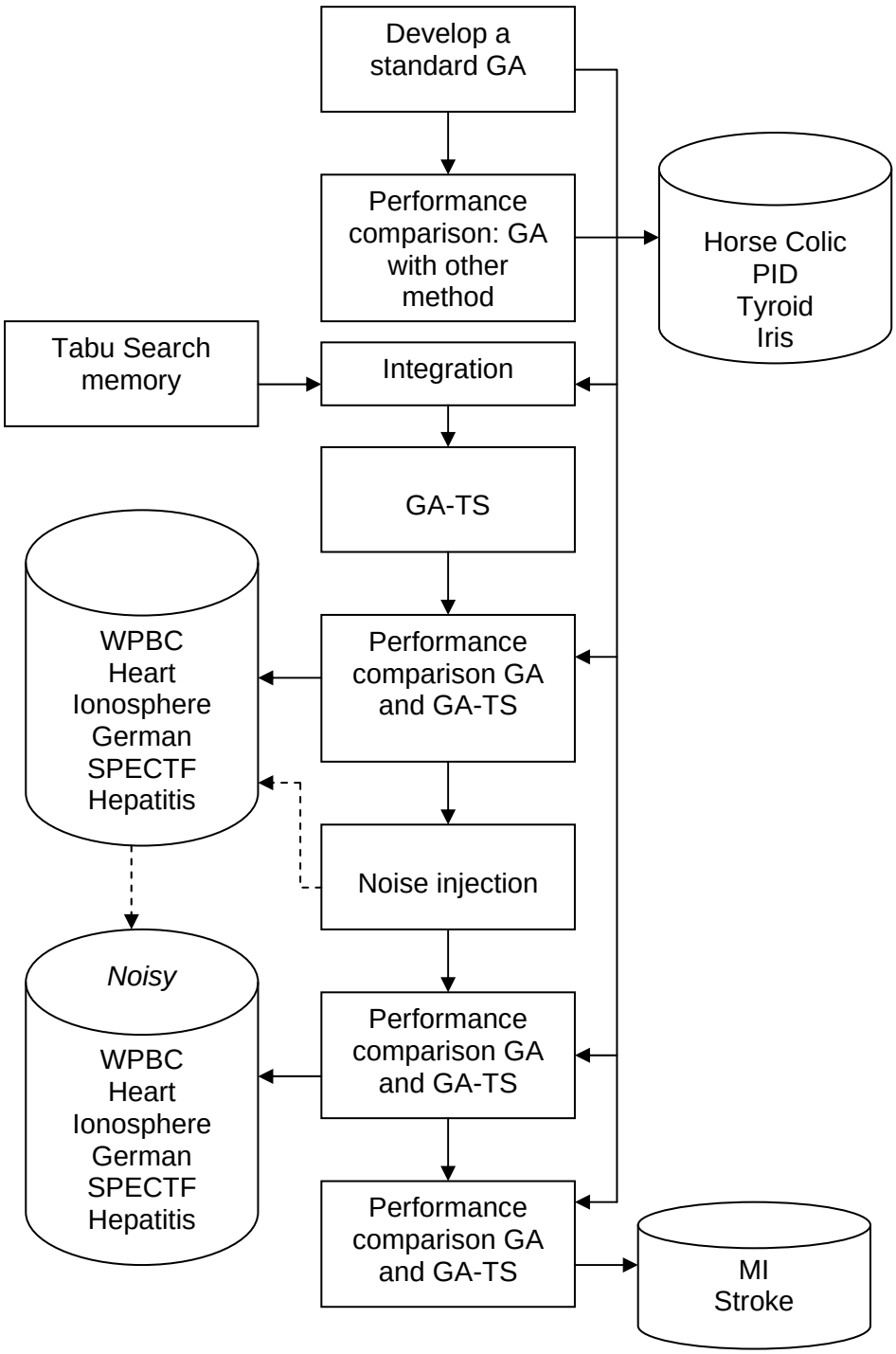

Figure 1.2 An overall diagram of research methodology applied in this research

1.9 Research Contribution

The applications of the GA in feature selection for FAM classification have been systematically investigated. The main contribution of this research lies in the proposal of an approach for feature selection using hybridisation of the GA and TS. Two memory structures, i.e. the recency and frequency memory structures, from TS are integrated into the genetic operators of the GA. To accommodate the TS memory structures in the GA, a new statistical rule is introduced to keep track of the frequency memory within the evolution process of the GA. The effectiveness of the proposed GA-TS algorithm coupled with FAM is systematically evaluated using a series empirical studies comprising benchmark data sets, and the results are analysed and compared with those published in the literature. In addition, the applicability of the proposed approach to medical pattern classification problems is demonstrated using real data sets collected from hospitals. In summary, this research proposes a new automated feature selection algorithm based on GA and TS for NN-based classification systems.

1.10 Thesis Outline

This thesis is organized in accordance with the objectives mentioned above. After an introduction to the research background in pattern classification and NN models, as well as an explanation to problems and motivations of the research, the research objectives, scope and methodology in Chapter 1, a general review on various search algorithms is presented in Chapter 2.

In Chapter 3, a review on GA is given. A series of experimental studies to evaluate the suitability of GA in feature selection are also presented. In Chapter 4, a review on TS is presented. Then, a hybrid GA-TS algorithm is proposed. A series of benchmark studies to evaluate the effectiveness of the proposed hybrid system are then presented.

The capabilities of the proposed hybrid system in noisy feature reduction are evaluated in chapter 5. Application of the proposed system in real-world case studies are conducted in Chapter 6. The results obtained are analyzed, compared, and discussed.

Finally, conclusions are drawn and contributions of this research are set out in Chapter 7. A number of areas to be pursued as further work are suggested at the end of this thesis.

CHAPTER 2

A REVIEW ON FEATURE SELECTION FOR NEURAL-NETWORK-BASED CLASSIFICATION

2.1 Introduction

Neural Networks (NNs) is one of the branches of Artificial Intelligence (AI). It is also known as connectionist systems, parallel distributed processing (PDP), neural computing, and artificial neural systems. NNs try to mimic how a brain and nervous system works. NNs are famous for their inductive learning ability, i.e., the ability of learning from examples. Besides that, NNs are able to generalize and recognize previously unseen data. NNs extract information without specifying a data model and possess the *train-and-go* characteristic.

The next section discusses the importance of feature selection in NN models. A classification of feature selection techniques is presented. A thorough survey on search techniques, which constitute the main strategy used for feature selection and for optimization purposes, is included. The operation of each search method is described. The advantages and disadvantages of the reviewed methods are also discussed. Then, a review on use of the Genetic Algorithm (GA) in feature selection is presented. A summary is presented at the end of this chapter.

2.2 Importance of Feature Selection in NN Models

In NN-based classification systems, disregard of the underlying NN model used, the classification performance is mainly dependent on the data set. The variation of the learning methodology in NN models only helps boost the classification performance. Therefore, the task of selecting useful and meaningful data for NN-based classification

is utmost important. In real-world situations, relevant features in a data set are often unknown *a priori* (Dash and Liu, 1997). In such situations, many features are introduced with the assumption that all features can help represent the domain problem. However, many of these features may be partially, if not completely, irrelevant to the domain problem. The existence of such redundant features can cause confusion during the learning phase, and also can increase the complexity of the feature space, which will result in a greater computational demand (Muni *et al*, 2006). Therefore, feature selection is a very important topic in NN-based classification.

Feature selection has caught the attention of researchers for quite some time. With the creation of new databases and new machine learning techniques, novel approaches for feature selection is in demand (Dash and Liu, 1997). Parkins and Nandi (2005) implemented feature selection in handwritten digit recognition while Garret *et al* applied feature selection to MLP for EEG signal classification. Mahil and Gao (2004) used feature selection on feedforward NN and RBF NN for machine defect classification problem. Jack and Nandi (2000) applied a GA in feature selection of vibration signals for machine condition monitoring problems with the MLP NN. Although a variety of feature selection methods have been introduced, all feature selection approaches share some common goals (Steppe, 1998):

- maximizing the classification accuracy while minimizing the number of features
- improving the classification accuracy by removing irrelevant features
- reducing the data complexity and computation cost
- reducing the amount of data for the learning phase
- improving the changes that a solution will both be understandable and practical

2.3 A Classification of Feature Selection Techniques

Generally, there are two categories of feature selection techniques (Law *et al* 2004, Blum and Langley 1997, Kohavi and John 1997): *filters* and *wrappers.* In filter

approaches, the data set alone is used to evaluate the relevance of each feature to the target output, regardless of the classification algorithm. Some representative algorithms in filters approaches are RELIEF (Kira and Rendell 1992) and its enhancement (Kononenko, 1994). The basic concept of these algorithms is to assign feature weights based on the consistency of the feature value in *k*-nearest neighbours of every data point. An information-theoretic method is used in Battiti (1994) to evaluate the features, whereby the mutual information between a relevant feature and the class label should be high. Besides, the concept of Markov blanket is used to formalize the notion of irrelevancy (Koller and Sahami, 1996); a feature can be regarded as irrelevant if it is conditionally independent of class labels given other features.

In wrapper approaches (Chaturvedi and Carroll, 1997), learning algorithms are used to evaluate the quality of each feature. Specifically, a learning algorithm is run on a feature subset, and the classification accuracy of the feature subset is taken as a measure for feature quality. Generally, wrapper approaches are more computational demanding as compared with filter approaches. However, wrapper approaches often are superior in accuracy when compared with filters approaches which ignore the properties of the learning task in hand (Chaturvedi and Carroll 1997). In most application of NN classification tasks, accuracy plays a greater role as compared with that of computational cost. Therefore, this research focuses on wrapper approaches for feature selection.

Both approaches, filters and wrappers, usually involve combinatorial searches through the space of possible feature subsets. In Chaturvedi and Carroll (1997), Pudil *et al* (1994), Law *et al* (2004), Yang and Honavar (1998), various search methods including sequential forward/backward searches, floating search, beam search, bidirectional search, and genetic algorithm are applied to feature selection.

2.4 A Review on Search Techniques

The goal in optimization is to find an optimal arrangement, grouping, ordering, or selection of discrete objects normally finite in number (Lawler, 1976). The approaches to combinatorial optimization problems are exact algorithms & approximate algorithms. Exact algorithms systematically search the solution space to find an optimal solution in finite time. The results of an optimal solution is guaranteed, but the time needed to solve NP-hard of many combinatorial optimization problems may grow exponentially in the worst case. On the other hand, approximate algorithms aim to get good and approximately optimal solutions in a reasonable time. In contrast to exact algorithms, approximate algorithms cannot guarantee optimality of the solutions returned. However, approximate algorithms such as local search and solution construction algorithms proved to achieve short computational time.

2.4.1 Blind search

In traditional search algorithms, the search process can be defined as a form of search tree where a goal state is defined. A node in the tree represents a solution, and its successors, or children, are defined by the operators. Search strategies, such as *blind* search, are defined to search the tree. Blind search methods typically do not have information of the problem domain. It is able to distinguish a non-goal state from a goal state. The two main *blind* search strategies are *depth-first* (Tarjan, 1972;Skiena, 1990) and *breadth-first* (Skiena, 1990) searches.

Breadth-first search is a systematic search where it generates all the successors of the root node in the search tree. Next, it generates all the successors of those nodes. It considers all nodes at each level of the tree and continues until a solution is found. Thus, it is guaranteed to find a solution if the solution exist. On the other hand, in depth-first search, a single successor of the root node is generated. It

follows by generating one successor to that node and continues until a maximum depth is reached. It explores though one branch of the tree and backtracks to the previous root node and generates another successor if no solution is found. It does not guarantee to reach goal state and may not reach the optimal solution.

Blind search methods are costly if the problem domain has a large search space. In the worst case scenario, a search problem cannot be solved in polynomial time of the size of the problem instance. As such, if the decision problem is in the NP hard complexity class, the optimization or search problem must also be at least as hard (Papadimitriou and Steiglitz, 1982).

2.4.2 Heuristic search

Heuristic search is introduced to solve problems with a large search space. It is an approximate method, which uses rules-of-thumb to define the problem structure. Such definition allows the generation of a possible solution in a combinatorial optimization problem or a search strategy that finds good solutions in a reasonable time (Reeves, 1995). Heuristics can improve time complexity in search problems by only considering a subset of all possible solutions or by only generating solutions that are closest to the goal state. A move in heuristic search is often made by applying a small random change on the current solution. Such operator is known as *mutation*, and normally it does not require any domain knowledge.

Hill-climbing or *steepest descent* (Arfken,1985) is an example of heuristic search. It only keeps track on one current state and moves on to the path that leads closer to the goal state. It begins with a randomly generated initial state. It then takes the successors of the current state and uses the evaluation function to assign a score to each successor. The successor with a better score is then set as the new current state. This process repeats iteratively until no changes in the current state occur.

Local optima may be reached and may cause incomplete search. Hill climbing search works well if an accurate heuristic measure is available in the domain, and if there are no local maxima.

Several variants of hill-climbing are proposed, e.g. *Stochastic hill-climbing* (William, 1988) and *multi-start hill-climbing* (Torn and Zilinskas, 1989). Unlike normal hill-climbing, stochastic hill-climbing also accepts neighbors that are equivalent to the current solution. The multi-start hill-climbing method starts the search by several random initial states in its attempt to improve the search.

Another simple heuristic search method is *best-first* (Pearl, 1984). This method starts with generating a set of neighbor solution based on the current solution. The best solution of the neighbour is selected. Then, based one best solution found in the neighbor solution, a new set of neighbor solution is generated. This process continues until no improvement can be found. Beam search, another search technique, is similar to best-first, but the difference is beam search will return to previous neighbor if no improvement can be found.

Heuristic search is often known as *local search*, as it focuses on searching for improved solution within the local neighbourhood of the current solution. Domain knowledge is only required as an evaluation function in heuristic search to measure the distance between current solution with the desired goal. Based on the improvement within local neighbours, the search will continue its move towards the desired goal. The heuristic evaluation function is also known as cost function, objective function, or fitness function.