EVOLUTIONARY COMPUTATION FOR MODEL STRUCTURE SELECTION IN SYSTEM IDENTIFICATION

MD FAHMI BIN ABD SAMAD @ MAHMOOD

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy (Mechanical Engineering)

Faculty of Mechanical Engineering
Universiti Teknologi Malaysia

APRIL 2009

ACKNOWLEDGEMENTS

Praise be to Allah the Cherisher and Sustainer of the Worlds for all His blessings. My appreciation goes to Prof. Dr. Hishamuddin bin Jamaluddin, as my supervisor, for his relentless support and guidance in the completion of this thesis. Thanks to Dr. Robiah bte Ahmad and Associate Professor Dr. Mohd. Shafiek bin Yaacob, my co-supervisors, for whom supervision I am also indebted. All are from Universiti Teknologi Malaysia, Skudai. Next, thanks to Assistant Professor Dr. Abul K. M. Azad from Northern Illinois University, my external supervisor, for his advices throughout the research.

My gratitude to the management group of my working institution, Universiti Teknikal Malaysia Melaka, especially to Datuk Prof. Ir. Ismail bin Hassan, former Vice Chancellor and Prof. Dr. Ahmad Yusoff bin Hassan, current Vice Chancellor. Thanks for the financial support, through UTeM-SLAB, and the opportunity for this study. I'd also like to thank the staffs of Unit Cuti Belajar, Bahagian Pengurusan Organisasi & Sumber Manusia, UTeM, School of Graduate Studies, UTM, Faculty of Mechanical Engineering, UTM and also others who helped me in any way.

To my parents and family members, thanks for continuous motivation in the undertaking of this "journey". I wish the same for future "journeys", insyaAllah.

ABSTRACT

System identification is a field of study involving the derivation of a mathematical model to explain the dynamical behaviour of a system. One of the steps in system identification is model structure selection which involves the selection of variables and terms of a model. Several important criteria for a desirable model structure include its accuracy in future prediction and model parsimony. A parsimonious model structure is desirable in enabling easy control design. This research explores the use of Evolutionary Computation (EC) in model structure selection. The effectiveness of penalty function in the objective function of EC is investigated. The results show that a suitable penalty function parameter can be achieved by its relation to the smallest estimated and tolerable parameter value. Using this function, an algorithm named Modified Genetic Algorithm (MGA) is proposed as it is able to reduce the possibility of premature convergence. MGA is proven to be more efficient than the original genetic algorithm where it is able to find a parsimonious model within a fixed or even shorter evolution period. Another algorithm, named Deterministic Mutation Algorithm (DMA) is proposed to reduce computational burden and reliance on optimum algorithm parameter setting. DMA is a simpler procedure that is able to assist user to obtain a parsimonious model within a shorter time. All of these system identification techniques are carried out by applying the algorithms to a number of simulated and real-life systems, namely gas furnace, Wölfer sunspot and hairdryer, using discrete-time models. Validations of the model structures are made using correlation tests and cross-validation.

ABSTRAK

Pengenalpastian sistem merupakan satu bidang kajian bagi menerbitkan model matematik untuk memperihal kelakunan dinamik sesuatu sistem. Satu daripada langkah dalam pengenalpastian sistem adalah pemilihan struktur model yang mana melibatkan pemilihan pemboleh ubah dan sebutan bagi sesuatu model. Beberapa kriteria penting bagi struktur model yang diingini ialah kejituan ramalan masa hadapan dan keringkasan model. Struktur model yang ringkas diingini untuk membolehkan reka bentuk kawalan yang mudah. Penyelidikan ini menyingkap penggunaan Komputasi Evolusi dalam pemilihan struktur model. Keberkesanan rangkap denda dalam rangkap objektif Komputasi Evolusi diselidiki. Keputusan menunjukkan bahawa parameter rangkap denda yang sesuai boleh diperolehi melalui hubungannya dengan nilai parameter terkecil yang dianggar dan boleh diterima. Dengan menggunakan rangkap ini, satu algoritma, dinamakan Algoritma Genetik Ubahsuaian dicadangkan kerana ia boleh mengurangkan kemungkinan penumpuan pramatang. Algoritma Genetik Ubahsuaian ditunjukkan lebih cekap daripada algoritma genetik asal di mana ia boleh mencari model termudah dalam tempoh evolusi yang sama atau lebih singkat. Satu lagi algoritma, dinamakan Algoritma Mutasi Berketentuan, dicadangkan untuk mengurangkan beban komputasi dan kebergantungan kepada ketetapan parameter algoritma yang optimum. Algoritma Mutasi Berketentuan merupakan prosedur yang lebih mudah dan boleh membantu pengguna memilih model termudah dalam masa yang lebih singkat. Kesemua pengenalpastian sistem dijalankan dengan menggunakan algoritma ini ke atas beberapa sistem simulasi dan sistem sebenar, iaitu relau gas, tompok matahari Wölfer dan pengering rambut, menggunakan model masa-diskret. Pengesahan model dibuat dengan menggunakan ujian sekaitan dan pengesahan silang.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS/NOTATIONS

a_i, b_i and c_i	-	Parameter values of regressor i
Acc	-	Number of chromosomes in group Acceptable
C_t	-	Best chromosome at generation i
d	-	Time delay
d, f and g	-	Model-dependent variables
d. c.	-	Constant and standalone parameter value
$e(t)$	-	Noise value at time t
$E[\cdot]$	-	Expectation operator
$F_*^l[\cdot]$	-	Nonlinear function of l degree
f_i	-	Fitness value of individual i
$f(H)$	-	Fitness value of schema H
J	-	Number of trials
k (as in k-point)	-	Number of crossover point
k (as in k-step-ahead)	-	Smallest lag order of predicted output in a difference equation model
L	-	Maximum number of regressors
l	-	Degree of nonlinearity
lchrom	-	Length of bit string chromosome
L_t	-	Locus of significant bit 1 in chromosome C_t at generation i
M	-	A model set
maxgen	-	Maximum number of generations
$m(H, t)$	-	Number of chromosomes of schema H at time t
M*	-	A set of parsimonious models

N	-	Total number of data
n	-	Number of regressors with parameter values less than or equal to *penalty* added by 1
n_e	-	Maximum noise lag order
n_u	-	Maximum input lag order
n_y	-	Maximum output lag order
Ord	-	Number of chromosomes in group Ordinary
p, q and r	-	Model-dependent variables
parsimony penalty	-	Penalty parameter where two models of different number of regressors have the same OF value but different EI
penalty	-	Parameter in a penalty function
popsize	-	Population size
p_c	-	Crossover probability or rate
p_i	-	Selection probability of individual i
p_m	-	Mutation probability or rate
R^2	-	Multiple correlation coefficient squared
s	-	Standard deviation of x
switchover penalty	-	Penalty parameter where number of significant regressors is equal to number of insignificant regressors
$u(t)$	-	Input value at time t
V_N	-	Criterion or cost function considering N data
w	-	Connection weight of a neural network
x	-	Number of generations needed to discover a model that is the best model until *maxgen*
$y_{\max}$ and $y_{\min}$	-	Maximum and minimum values of output, respectively
$y(t)$	-	Output value at time t
$\hat{y}(t)$	-	Predicted output value at time t
Z^N	-	Function of parameter vector $\boldsymbol{\theta}$

$\delta(H)$	-	Defining length of schema H
$\delta(\tau)$	-	Kronecker delta
ε	-	Residual or prediction error
$\boldsymbol{\theta}$	-	True parameter vector
$\hat{\boldsymbol{\theta}}$	-	Estimated parameter vector
λ	-	Number of offsprings in evolution strategies
μ	-	Number of parents in evolution strategies
$o(H)$	-	Order of schema H
σ	-	Variable in a chromosome schema
τ	-	Lag order
ϕ_i	–	Standard correlation function of i
$\boldsymbol{\phi}$	-	Regressor vector
□	-	Wildcard symbol in a chromosome schema

LIST OF ABBREVIATIONS

AR	-	AutoRegressive
ARMAX	-	AutoRegressive Moving Average with eXogenous input
ARX	-	AutoRegresive with eXogenous input
CA	-	Cellular automaton (singular) or cellular automata (plural)
DMA	-	Deterministic mutation algorithm
EA	-	Evolutionary algorithms
EC	-	Evolutionary computation
EI	-	Error index
EP	-	Evolutionary programming
ERR	-	Error reduction ratio
ES	-	Evolution strategies
FF	-	Fitness function
GA	-	Genetic algorithm
GP	-	Genetic programming
HC	-	Hill-climbing
OF	-	Objective function
OLS	-	Orthogonal least squares
PSO	-	Particle swarm optimization
SGA	-	Simple genetic algorithm
SISO	-	Single-input-single-output
MGA	-	Modified genetic algorithm
NAR	-	Nonlinear AutoRegressive
NARMAX	-	Nonlinear AutoRegressive Moving Average with eXogenous input

NARX	-	Nonlinear AutoRegresive with eXogenous input
NOE	-	Nonlinear output order
PF	-	Penalty function

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Introduction

System identification is a method of determining a mathematical model for a system given a set of input-output data of the system (Johansson, 1993). There are four main steps involved in system identification and these are data acquisition, model structure selection, parameter estimation and model validation (Söderström and Stoica, 1989; Ljung, 1999). As one of the stage in system identification, the model structure selection stage refers to the determination of the variables and terms to be included in a model. Basically, an optimum model is described as having adequate predictive accuracy to the system response yet parsimonious in structure. A parsimonious model structure is preferred since, with less number of variables and/or terms, system analysis and control becomes easier.

Traditionally, model structure selection is performed by determining a finite set of models, typically within a certain maximum specification, and enumeratively testing the models for predictive accuracy and parsimony. The decision of selection is based on certain information criterion where some established criterions are Akaike's information criterion, B-information criterion and ϕ-information criterion (Veres, 1991). Another method reported is the regression methods such as the backward elimination, forward selection or inclusion and stepwise regression method. These methods involve testing of

different models guided by an analysis of each model's squared multiple correlation coefficient, R^2 and partial F-test value (Draper and Smith, 1998). In another development, a method called orthogonal least squares is applied in model structure selection (Korenberg *et al*., 1988; Billings and Yang, 2003a). Despite these encouraging developments, these methods require heavy statistical computation. In order to overcome this, researchers turn to search methods that are able to provide a selection method that is simpler and more efficient in term of cost and time.

The most recent and successful search method applied to system identification is evolutionary computation (EC) (Fleming and Purshouse, 2002). EC is a term known since 1991 to represent a cluster of methods that uses the metaphor of natural biological evolution in its search and optimization approach (Fogel, 2000). Unlike conventional search methods, EC searches from a global perspective i.e. it does not settle with a local optimum solution (Sarker *et al.*, 2002). Its search is guided by an evaluation function, also called objective function (OF), where good information is exploited via genetic operators. Generally, these operators are reproduction, crossover and mutation. This capability enables the determination of optimum solutions to various optimization problems.

The current research and development in evolutionary computation lists three major areas that are evolutionary computation theory, evolutionary optimization and evolutionary learning. Evolutionary optimization is mentioned to be the most active and productive area (Sarker *et al.*, 2002). EC applications are known in various fields, among others are power system optimization, control systems engineering and manufacturing optimization (Alves da Silva and Abrão, 2002; Fleming and Purshouse, 2002; Dimopoulos and Zalzala, 2000).

1.2 Problem Statement

Model structure selection in system identification basically involves the search for an optimum model structure among many alternative models. This can be achieved by using a search method. Conventional search methods, namely simulated annealing, tabu search and hill-climbing algorithm, have been applied for optimization problems. However, conventional algorithms conduct its search within a local landscape (Mitchell, 1996; Sarker *et al*., 2002; Michalewicz, 1996). Due to this, the methods have the tendency to converge to local optima, giving sub-optimal model structure to a system identification problem.

The characteristic of global search is found in EC where it is able to perform the search for an optimum model by exploiting good information via global manipulation of solutions. However, its ability is restricted when more efficient search is required especially when constraints like parsimony of model structure is present in the problem. Past researches usually concentrated on predictive accuracy and only few treated the issue of model parsimony, yet still with some inadequate justification (Ahmad *et al*, 2004a). In this regard, a more suitable objective function is needed. This can be found by an understanding of the relationship between certain specified OF to the result of model structure selection.

From another viewpoint, EC search is also disadvantageous as it needs cumbersome setup of user-defined parameters for the algorithm, referred as algorithm parameters in Eiben *et al.* (2007), and long computational time. Although the convergence of EC to global optimum is theoretically achievable with a modest setting, the most efficient algorithm should converge with the simplest or optimum setting of the parameters. These algorithm parameters include population size, number of generation, representation, crossover type, mutation type, probability of crossover, probability of mutation and mating strategy (Bäck *et al.*, 2000a; 2000b).

A limitation of EC that is related to poor setting of its parameters is premature convergence. This happens when the best few members of a population in the algorithm predominate the population. In short, an issue that needs addressing is not only developing an algorithm that has good convergence properties but also assuring that it converges in the direction of the global optimum solution. Although other techniques have been applied to overcome this problem, an imbalance to other priorities seems to arise. For example, certain selection methods help in increasing diversification of population but at the expense of a longer search time. A method that reduces or overcomes these limitations is thus needed. Among strategies that seem feasible in achieving this is through a re-evaluation of objective function in EC and modification of the procedure, especially by the elimination of the factors that contribute to the weaknesses.

1.3 Research Objectives

Several objectives are identified for this research and these are stated and explained as follows:

(i) To propose an alternative algorithm for model structure selection that overcomes the limitations of conventional algorithms and evolutionary computation.

The proposal of an alternative algorithm is mainly based on genetic algorithm, which is the most well-known algorithm in EC. The purpose of the alternative algorithm is to be used for the determination of variables and terms to be included during model structure selection. During the development of the algorithm, several issues that arise are global search capability, probability of premature convergence, algorithm setup, computational complexity and effectiveness of solution in term of adequacy and parsimony. Among questions to be answered are ‘What are

the right setting of parameters for the search?' and 'What evaluation function should be used?'

(ii) To show that the algorithm is applicable.

The applicability of the algorithm is to be shown using simulated data modelled by the user. Simulation studies are beneficial because the studies enable direct comparison of selected model structures by the algorithm to the correct ones. Disturbances are also purposely injected to the models to resemble realistic situation. In the final stage of system identification, validation is performed to verify the adequacy of the model.

(iii) To model real-life problems those are widely discussed in academic circle.

The performance of the algorithm is further evaluated by implementing it to real-life modelling problems. Problems that are present in literature provide direct benchmarking opportunity in the study. Some real-life problems that are available in literature include the Wölfer sunspot time series data and gas furnace data (Box *et al.,* 1994; Jenkins and Watts, 1968). The Wölfer sunspot data is an example of a one-variable time series data where no input is present, while the gas furnace data is a single-input-single-output (SISO) data. Lastly, an internet database of real-life raw data, called DaISy: Database for the Identification of Systems, provides another source for testing real-life problems like a hairdryer system (De Moor, 2008).

1.4 Research Scopes

Due to wide development of study in the field of system identification, the research is limited to the following scopes:

(i) Only discrete-time difference equation models were used.
With the assumption that the output of a system is a realization of the variables at instants of time, discrete time models (also called time series model) become a practical choice. The assumption is also inline with typical data acquisition practice. In the group of discrete-time models, difference equation model is the simplest interpretation of a system's process. A study of difference equation models has shown that difference equation models are representative of many other types of models (Chen and Billings, 1989). A common linear model structure for discrete-time systems is the ARX (AutoRegressive with eXogenous input) model. A nonlinear ARX (NARX) model is used to represent a nonlinear discrete-time system.

(ii) Data consisted of less than two input and/or output variables.
The testing of the algorithm was made on data those are in the form of single input-single output and time series. It does not, however, restrict its applicability to data of more than two variables since the application of EC to this type of data only requires minor rearrangement of data and is not considered as a new subject (Ahmad *et al.*, 2002).

(iii) The least squares method was used for estimation of system parameters.
For simulated models, the disturbances were injected from a uniform distribution. In this circumstance, the least squares method becomes an unbiased method since the disturbances infinitesimally behave as white noise. This form of disturbances also suggests that the noise data are uncorrelated which is suitable for the least squares method. The least squares method also becomes a generalization to other methods like maximum likelihood (Draper and Smith, 1998). The assumption of white noise is also used for real-life problems. The method is widely used in literature and the simplest when the assumption is true.

(iv) Comparisons of research findings were made to literature findings and similar methods.

When comparing the performance of an alternative algorithm, only findings from literature and similar methods were used. No statistical method is redo for comparison. Furthermore, a comparison of a modified genetic algorithm has been shown to be equally good or better than a statistical method that is considered popular, today – the orthogonal least squares (Ahmad *et al.*, 2004a; 2004b).

1.5 Research Methodology

The methodology of the research is based on the general flow of system identification which includes data acquisition, model structure selection, parameter estimation and model validation, as shown in Figure 1.1. Although the main purpose of the research is to propose an alternative algorithm for model structure selection, the research also considers other aspects of the flow. Every stage is defined and carried out so that the standard procedure of system identification is clearly accomplished and the applicability of the whole proposal is clarified.

The development of the alternative algorithm is related directly to the model structure selection step. The step is broken down into several other steps as follows:

(i) Identifying and understanding the weaknesses/inadequacies in established methods.

(ii) Developing a method that overcomes the weaknesses/inadequacies by modifying/renewing the procedure of an established method.

(iii) Evaluating the performance of the developed method among its own variants or other original methods.

(iv) Repeating steps (i) to (iii) for further development of the developed method until a satisfactory algorithm is established.

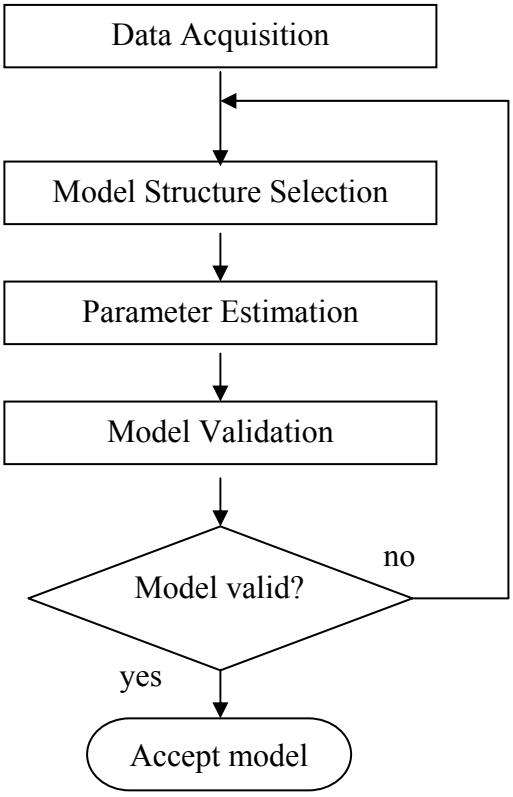

Figure 1.1 Flow chart of research methodology

The flow chart of the steps for algorithm development is provided in Figure 1.2. It has to be noted here that several weaknesses are present in EC, as provided in Section 1.2, even by considering only EC methods that are developed for model structure selection. Due to this reason, the first three steps above are repeated until an algorithm that is more superior than its original method is established. Although one might choose to see this methodology as a continuous flow by keep modifying the algorithm, it is presented here as ending with a final algorithm within the time-frame of the research.

With regards to the comparison of algorithms during the testing on simulated and real-life problems, several common performance indicators are used such as predictive accuracy, model parsimony and computation time. Besides these measures, results are also compared to literature findings and via validation methods like correlation tests and *k*-step-ahead simulation.

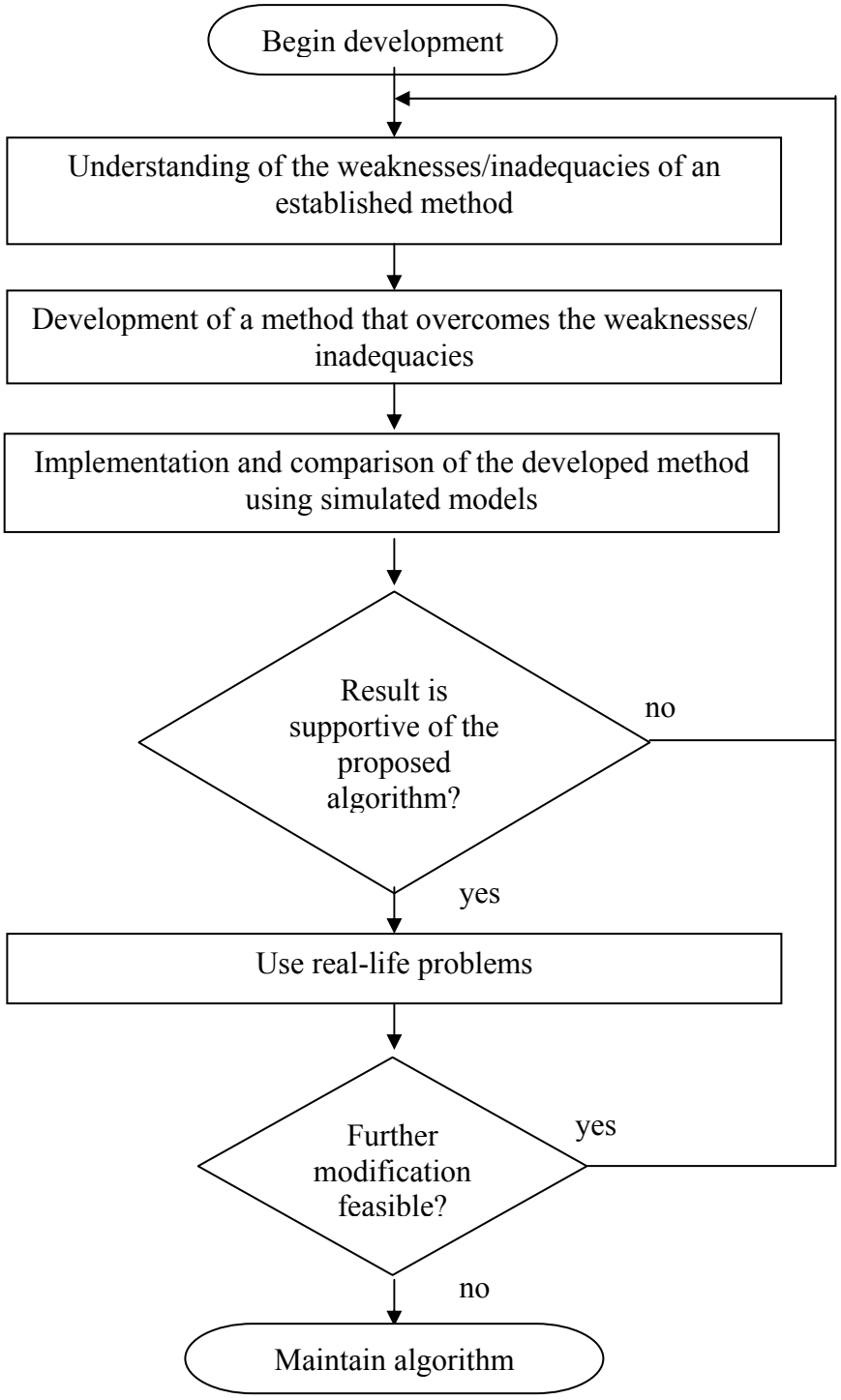

Figure 1.2 Flow chart of alternative algorithm development

1.6 Research Contributions

The aim of the research is to propose an alternative algorithm for use in model structure selection for system identification. Before any new algorithm is proposed and compared to other search methods, the effectiveness of the objective function (OF) as used in Ahmad *et al.* (2004a) is investigated. The first contribution of the research is the provision of a clear relationship between the selected OF and the result of model structure selection. A guide on the selection of a suitable penalty parameter that provides an adequate and parsimonious model is also presented.

The second contribution of the research revolves around the proposal of a modified genetic algorithm (MGA). The idea behind the modification is through grouping of population for different manipulation. Although the idea have been used in Ahmad *et al.* (2004b), the implementation was rather case-based. This research provides a more clear-cut method of how the grouping should be done.

The last contribution is the proposal of another algorithm, named deterministic mutation algorithm (DMA). This algorithm takes advantage of the implicit parallelism theory as defined by Holland (1992). The introduction of 'wildcard attribute' in the theory is exploited for model structure selection problem and combined with an element of forward search. The strengths of the algorithm are its reduction of the reliance for optimum algorithm setting, better parsimonious model search and less computation time.

1.7 Organization of the Thesis

This thesis comprises of six chapters. The first chapter introduces the background of the research. This is followed by an explanation of the problem to be tackled. The objectives and scopes of the research are then laid out and the research

methodology is described. A brief explanation of the organization of the thesis is also provided.

The second chapter reviews various literature related to the study mainly on system identification and evolutionary computation. In the early part of the chapter, the steps of system identification are explained. This explanation covers various choices of model types, considerations in constructing an optimum model structure and methods of implementing parameter estimation and model validation. Next, several methods applied for model structure selection are explained along with some identified disadvantages. The later part of the chapter discusses EC and its four specific methods – genetic algorithm, evolution strategies, evolutionary programming and genetic programming. Examples of EC application in modelling are given. Next, the chapter reviews recent EC literature on the aspect of algorithm procedures for system identification followed by explanations of some common procedures. Potential areas for research are provided at the end before the summary of the chapter.

The third chapter deals with an investigation of the suitability of an objective function for model structure selection. The chapter begins with an explanation of NARX model structure representation and the least squares method as its parameter estimation method. Then, genetic algorithm as its search method is explained in terms of its procedure, theoretical foundation and other related aspects. This is followed by a background of the study where a logarithmic penalty function with a penalty parameter is tested on five simulated models. The discussion of the results is supplemented with visual presentation of the relationship between the OF and the results of model structure selection. A discussion on the selection of a suitable penalty parameter is given. The shortcomings of the method are also provided.

Chapter 4 explains a modified genetic algorithm (MGA) that stresses on grouping of the solution population by a fixed ratio. Two groups and two individuals of different fitness values are manipulated differently. A discussion on a model validation method based on correlation tests is also presented. Based on the tests on two simulated

models and two real-life problems, a variant of MGA denoted Ratio 3, is proven to produce more accurate model structure or requires less generation in producing the same model structure compared to other variants. One of the other variant is similar to a simple genetic algorithm.

Another alternative algorithm, called deterministic mutation algorithm (DMA), is explained in Chapter 5. The theoretical foundation and procedure of DMA is provided where, among others, explains its contribution in escaping from the usual reliance of evolutionary computation on algorithm setting. Its differences to hill-climbing algorithms are also given. The background of the simulation study are given along with an explanation of the cross-validation method. Three simulated models and three real-life problems are tested and the results show that DMA has the advantage as a model structure selection method that easily balances accuracy and model parsimony and requires shorter computation time.

The last chapter recaps the application of evolutionary computation in model structure selection and its downfalls. It lists the findings of the research, namely in the usage of penalty function in objective function and the performance of the algorithms – modified genetic algorithm and deterministic mutation algorithm. Several recommendations for future research directions are also given.