IMPROVED GENETIC ALGORITHM – MULTILAYER PERCEPTRON NETWORK FOR DATA CLASSIFICATION

FADZIL BIN AHMAD

UNIVERSITI SAINS MALAYSIA

2017

IMPROVED GENETIC ALGORITHM – MULTILAYER PERCEPTRON NETWORK FOR DATA CLASSIFICATION

by

FADZIL BIN AHMAD

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

January 2017

ACKNOWLEDGEMENT

"In the name of ALLAH, Most Gracious, Most Merciful"

All praises to Allah, we seek His help and ask for His forgiveness. We seek refuge with Allah from the accursed *syaitan* and from evils of our souls and our bad deeds. I would like to give glory to Allah almighty for making it possible for me to complete this thesis and for His help in all my endeavors and for bringing me this far in my educational life.

I would like to express my deep gratitude and heartfelt thanks to my supervisor, Prof. Dr. Nor Ashidi Mat Isa, for his calmness, patience and creative guidance throughout this research work until the completion of this thesis. His intellectual supports and constructive criticisms greatly enhance the contribution of this thesis.

Great thanks from my heart to my lovely wife for her patience and support, my sons and daughter for their patience too, and my father-in-law, mother-in-law, brothers, sisters and friends for their encouragements. Not to forget the memory of my late mother and father for encouragement.

Finally, I wish to thank all lecturers, administration, academic and non-academic staffs of School of Electrical and Electronics Engineering, Universiti Sains Malaysia for their help and support. I would like to appreciate Universiti Sains Malaysia for funding this work through research grant entitled "Genetic Algorithm-Artificial Neural Network Hybrid Intelligence".

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACO	Ant Colony Optimization
ANN	Artificial Neural Network
BANI	Best Among Normal and Improved Population Elitism
BSSR	Best Between Similar Ranking Population Elitism
EC	Evolutionary Computational
ECG	Electrocardiography
EQ	Equally Contributed Elitism
FL	Fuzzy Logic System
FN	False Negative
FP	False Positive
GA	Genetic Algorithm
MLP	Multilayer Perceptron Network
MSE	Mean Squared Error
PSO	Particle Swarm Optimization
SI	Swarm Intelligent
SMCC	Segmented Multi Chromosome Crossover
STD	Standard
SVM	Support Vector Machine
TP	True Positive
TN	True Negative

PENAMBAHBAIKAN ALGORITMA GENETIK - RANGKAIAN PERSEPTRON BERBILANG LAPISAN UNTUK PENGKELASAN DATA

ABSTRAK

Secara umumnya, algoritma genetik (GA) konvensional mempunyai beberapa kelemahan seperti penumpuan pramatang, kecenderungan terperangkap pada penyelesaian optima setempat dan ketidakupayaan penalaan di sekitar kawasan berpotensi. Oleh itu, GA ditambahbaik dengan strategi pencarian, penghasilan semula dan elitisma baharu dicadangkan dalam kajian ini. Pernambahbaikan pertama melibatkan perubahan kepada struktur operasi GA yang mana ia menumpu pencarian di sekitar kawasan berpotensi tinggi. Kedua, teknik baharu penghasilan semula yang dinamakan *Segmented Multi-Chromosome Crossover* (SMCC) telah diperkenalkan. Teknik tersebut mengelak kemusnahan maklumat hampir optima yang terkandung dalam segmen genetik dan membolehkan generasi baharu mewarisi maklumat penting daripada berbilang induk. Ketiganya, tiga jenis variasi elitism dinamakan sebagai *Best Among Normal and Improved Population* (BANI), *Best Between Similar Rank* (BBSR) dan *Equally Contributed* (EQ) telah dibangunkan. Ia melibatkan pertandingan di kalangan individu terbaik daripada populasi normal dan ditambahbaik untuk kelangsungan pada generasi selepasnya. GA yang ditambahbaik kemudiannya digunakan untuk mengoptimasi dan merekabentuk rangkaian perseptron berbilang lapisan (MLP) secara automatik bagi penyelesaian masalah pengkelasan corak. Bilangan nod terlindung, nilai pemberat sambungan awalan dan pemilihan ciri MLP yang memainkan peranan penting dalam menentukan prestasi pengkelasan dipilih untuk dioptimasi secara automatik oleh GA ditambahbaik. Prestasi GA ditambahbaik telah dinilai menggunakan fungsi ujian penanda aras yang rumit serta berbilang mod dan dibandingkan dengan GA piawai. Berdasarkan kekerapan sesuatu algoritma menghasilkan keputusan terbaik terhadap fungsi ujian yang berbeza; ianya telah terbukti bahawa prestasi teknik yang dicadangkan mengatasi GA piawai. BANI, BBSR dan EQ mencatatkan 30, 18 dan 17 kekerapan keputusan terbaik masing-masing berbanding GA piawai yang hanya mencatatkan 3 keputusan terbaik. Manakala, prestasi pengkelasan GA-MLP yang ditambahbaik telah dinilai menggunakan set-set data yang berbeza dari segi saiz ciri kemasukan dan bilangan kelas keluaran. Keputusan menunjukkan keberkesanan algoritma baharu daripada segi peratusan ujian kejituan. Peratus peningkatan keseluruhan sebanyak 0.6%, 0.1% dan 0.3% bagi ujian kejituan dicatatkan oleh BANI, BBSR dan EQ berbanding dengan GA-MLP piawai.

IMPROVED GENETIC ALGORITHM-MULTILAYER PERCEPTRON NETWORK FOR DATA CLASSIFICATION

ABSTRACT

In general, conventional genetic algorithm (GA) has several drawbacks such as premature convergence, high tendency to get trapped in local optima solution and incapable of fine tuning around potential region. Thus, new improved GA that focuses on new search, reproduction and elitism strategy is proposed in this study. The first improvement involves changes in the operational structure of GA in which it concentrates the search in highly potential area in the search region. Secondly, a novel reproduction technique called *Segmented Multi-Chromosome Crossover* (SMCC) is introduced. The proposed technique avoids the destruction of nearly optimal information contained in the gene segment and allows offspring to inherit highly important information among multiple parents. Thirdly, three new variations of elitism scheme namely *Best Among Normal and Improved Population* (BANI), *Best Between Similar Rank* (BBSR) and *Equally Contributed* (EQ) are developed. It involves competition among best individuals from normal and improved population to ensure survival in the next generation. The improved GA is then applied for optimization and automatic design of multilayer perceptron (MLP) neural network in solving pattern classification problem. Hidden node size, initial weights and feature selection of the MLP that play significant role in the classification performance are selected to be automatically optimized by the improved GA. The performance of improved GA has been evaluated using highly complicated and multimodal benchmark test functions and compared with the standard GA. Based on the occurrences of the best result obtained by an algorithm across different test

functions; it is proven that the proposed method outperforms standard GA. BANI, BBSR and EQ scores 30, 18 and 17 occurrences respectively compared to the standard GA that only scores 3 occurrences. Meanwhile, the improved GA-MLP classification performance has been evaluated using datasets that vary in input features and output sizes. The results demonstrate the effectiveness of the new algorithms in term of test accuracy percentage. There is an overall improvement of 0.6%, 0.1% and 0.3% in test accuracy of BANI, BBSR and EQ compared to the standard GA-MLP.

CHAPTER 1

INTRODUCTION

1.1 Background

The general research area of this study is artificial intelligence that has significant contribution to the life style of modern civilization. A variety of new technologies and applications emerge resulted from the research advancement in this area. These include machines that learn and decide on their own actions (Galindo and Saffiotti, 2013), web crawlers that systematically index information in websites (Xu et al., 2014) and intelligent assistants that automatically detect financial fraud (Barraclough et al., 2013) or diagnose diseases (Nahar et al., 2013).

The term artificial intelligence was first introduced in 1956 by John McCarthy, during the first academic conference on the subject at Dartmouth College in New Hampshire (McCarthy et al., 2006). He defines artificial intelligence as

"*the science and engineering of making intelligent machines, especially intelligent computer programs*".

Another well accepted definition of artificial intelligence is provided by Bellman (1978), i.e.,

> *"the automation of activities that we associate with human thinking, activities such as decision-making, problem solving, learning".*

Meanwhile, Schalkoff (1990) defines artificial intelligence as

"*A field of study that seeks to explain and emulate intelligent behavior in terms of computational processes*"

In the perspective of engineering discipline, artificial intelligence deals with the development and building of intelligent machine and software that can perform activities requiring human intelligence. While the goal of pure science discipline emphasizes on understanding the intelligent behavior, engineering discipline focuses on the design, development and implementation of the intelligent system itself, together with the relevant concepts, theories, techniques and the applications. Tasks involve intelligent action include problem solving, classification, perception, optimization, learning, motion planning, natural language processing and etc. In order to design and develop a machine that exhibits these intelligent behaviors, various issues have to be first understood and addressed. For example, how machines can acquire, represent and store knowledge, and consequently learn and generate intelligent behaviour (Gruber, 2013). Another issue is on how to transform sensory signal into symbols to be manipulated by machines in order to perform logic operations (Akce et al., 2013). Due to the variety of issues to be addressed pertaining to the intelligent behavior of the machine, the research activities of artificial intelligence has branched out into a number of sub-areas such as knowledge representation, optimization, search, machine learning, pattern recognition and computational intelligence tools.

Pattern recognition (Cherkassky et al., 2012, Meyer-Baese and Schmid, 2014) continues to be of great interest among all the sub-areas of artificial intelligence, due to the rapid growth of applications that can benefit from it. It is an essential part of our daily routine that occurs without the conscious effort of the individual. Patterns are transferred to human brain via sensing organ to be processed into useful information, and consequently, the judgments and decision for the pattern are made (Duda et al., 2000). But, to computerize this system is not an easy task. The development of pattern recognition system involves the invention and integration of specific methods, algorithms and equipments in order to perform the task. Some of the important application areas of pattern recognitions are image and speech analysis (El Ayadi et al., 2011, Loo et al., 2014, Lopez-de-Ipina et al., 2014), character recognition (Li et al., 2012, Chacko et al., 2012), medical diagnostics (Foster et al., 2014, Yoo et al., 2014) and machine diagnostics (Najafi et al., 2012, Wang et al., 2011, Widodo et al., 2011), business forecasting and prediction (Hung and Lin, 2013, Moon et al., 2013, Nagar and Malone, 2011), person identification (Bauml et al., 2013, Langdon et al., 2014) and industrial inspection (Alarcon-Herrera et al., 2014, Nikolic et al., 2013).

In general, pattern recognition aims to provide acceptable answers corresponding to all possible sets of input pattern values. Classification is a typical example of pattern recognition. It is a process of assigning information, objects, or, in general, input patterns into a given set of classes. For example, determines whether a tissue sample is "cancerous" or "non-cancerous. Regression is another category of pattern recognition, in which a real output value, is assigned to the input set (i.e. stock market prediction). Other examples are parsing in which a parse tree is assigned to an input sentence and sequence labelling in which a class is assigned to a sequence of values (i.e. speech tagging).

In general, the steps involved in the design of pattern recognition system are described in Figure 1.1. Its implementation is explained as follows :

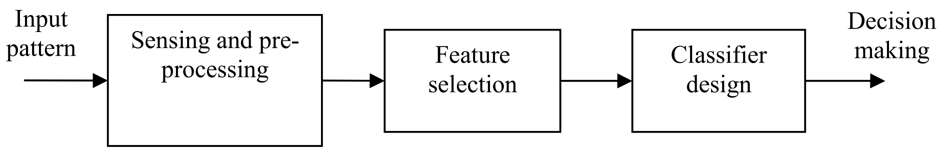

Figure 1.1 A general pattern recognition system

1) Sensing and pre-processing - design specification and requirement for this stage may vary on different application. For example, sensing equipment for image classification is different compared to those required for disease diagnosis. In some cases, pre-processing such as noise filtering and image segmentation are required in order to improve the data quality and to partition image into meaningful regions respectively.

2) Feature selection – selection of the significant features from the processed data using an arbitrary function. The collective combination of these relevant features represents the object to be recognized by the system.

3) Classifier design - the preferred operation of classification or recognition is designed and based upon the attributes of the selected features the appropriate output class or value is assigned to the input pattern.

Feature selection, also called variable subset selection or attribute selection, is an important component in designing a pattern recognition system. It is related to the selection of relevant and significant feature subset from full original feature set in designing the classifier. The motivation in implementing feature selection is mainly due to the assumption that the original data also contains irrelevant and redundant attributes. The benefits of feature selection include improving the classifier performance, providing less cost and faster classifier, and better understanding of the problem domain (Guyon and Elisseeff, 2003).

1.2 Approaches to Pattern Classification Problem

Numerous techniques have been developed in solving pattern classification problem. Among the earliest effort is the statistical based approach, the linear discrimination method suggested by Radhakrishna Roa (1948) and Fisher (1936). Another well-known statistical method for classification is Bayesian rule, proposed by Devijver and Kittler (1982). However, Pal and Pal (2001) argued that statistical methods are ineffective when dealing with contextual or structural data in patterns. As a solution, many works have switched to the theory of formal languages (Hopcroft et al., 2001), which made the syntactic approaches for pattern classification become popular. Patterns to be classified in syntactic approaches are not in a form of arrays of numbers. They are represented by simple sub-elements known as primitives, and associated with certain rules known as syntactical rules. However, these approaches have disadvantages of being ineffective in dealing with noisy and distorted patterns (Pal and Pal, 2001).

Classification tree (Gelfand et al., 1989, Lawrence et al., 2004) is another famous approach for pattern classification. It is a symbolic system that associates symbolic decisions to input examples, and is constructed using attributes of the examples that are symbolic in nature. A tree-like graph of decisions and their possible consequences are used to support the classification process. Nevertheless this classifier suffers similar inefficiency as syntactical approach when dealing with noisy and distorted environment (Pal and Pal, 2001).

A different powerful approach to pattern classification is based on artificial intelligence tools, or more specifically the computational intelligence models. It is a study of adaptive mechanisms to enable or facilitate intelligent behaviour in complex and changing environments (Engelbrecht, 2002). The computational intelligence models include four paradigms, i.e., Artificial Neural Network (ANN), Evolutionary

Computing (EC), Swamp Intelligence (SI) and Fuzzy Logic System (FL). Each of the computational intelligence paradigms has its origins in biological systems. ANN replicates human biological neural systems, EC models natural evolution of living population, SI models the social behaviour of organisms living in swarms or colonies, and FL originates from studies of how organisms interact with their environment. In fact, these computational models are powerful approaches for tackling pattern classification problem as demonstrated by the success in many new applications. For example, ANN has been successfully applied in the classification of brain tumour (Sridhar and Murali Krishna, 2013), partial discharge pattern in fingerprints input system (Abubakar Mas'ud et al., 2014) and bearing faults (Barakat et al., 2013). Recently, the SI approach, Ant Colony (ACO) and Particle Swarm Optimization (PSO) have been observed in classifying power signal (Biswal et al., 2011) and autism symptom (Oikonomou and Papageorgiou, 2013) respectively. Meanwhile it has also been observed that the EC variation, particularly the Genetic Algorithm (GA) has been applied for classification of Electrocardiography (ECG) signal (Martis et al., 2014) and subtypes of acute lymphoblastic leukemia (Lin et al., 2013).

Recent research trend in artificial intelligence tools for pattern classification has directed towards hybridization of computational intelligence models. The hybrid computational intelligence models utilize more than one problem-solving technique in order to solve a problem. Individual technique such as ANN, FL and GA are combined together to form a better intelligent system. This is achieved by exploiting the advantages of computational intelligence models involved in combination and avoiding its shortcomings as evidence by the success of many previous research efforts. Pattern classification approach that involves combination of EC and ANN can be observed in Fernandez et al. (2010), Mantzaris et al. (2011), Quteishat et al.

(2010) and Yang et al (2012). While in Alcala-Fdez et al.(2011), Aydogan et al. (2012) and Martin et al.(2011), the researches involve the combination of EC and FL.

Feature selection is another important approach to be considered in handling pattern classification problem. The objective of feature selection is to select a subset of features in order to enhance the classification accuracy or to reduce the size and complexity size of the classifier without significantly decrease its classification performance built using only the selected features (Dash and Liu, 1997). The simplest approach is by evaluating each possible feature subset while finding the best subset which gives highest classification rate. This approach however, is computationally expensive and impractical especially for high dimensional input data. Other approaches that use heuristic search methods such as GA and ACO attempt to minimize the computational cost and at the same time improve or maintain the model's classification performance.

1.3 Hybrid Genetic Algorithm-Artificial Neural Network

Since the last decades, numerous algorithms based on hybrid computational intelligence model have been developed for tackling pattern classification problems. This multidisciplinary research area continues to expand and become popular among the artificial intelligence research community. Hybrid computational intelligence models are defined as combination of computational intelligence paradigms and techniques that theoretically and practically fit as a basis for working in more competitive approach compared to single computational intelligence models (Corchado et al., 2010). They were developed either by integrating two or more computational intelligence paradigms, which preserves the characteristics of each technique, or by fusing one computational intelligence paradigm into another or by