Non-Linear Features and Multi-Objective Based Feature Selection Algorithms for Infant Cry Classification

by

LIM WEI JER
(1540611818)

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

School of Mechatronic Engineering
UNIVERSITI MALAYSIA PERLIS

2019

ACKNOWLEDGMENT

I would like to thank my supervisor Dr. Vikneswaran Vijean for his continuous encouragement, extensive support, freedom and trust that enabled this research to run smoothly. I am greatly indebted to him without whose invaluable guidance and encouragement this report could not have been completed

I would also like to express my gratitude to my co-supervisor, Dr. Haniza Yazid and Dr. Hariharan Muthusamy whose support, encouragement, and stimulating suggestions helped me in all the time during preparation for this report. They always showed a keen interest in the research process and while encouraging me to explore new ideas freely.

Last but not least, I would like to thank my family for the sacrifices they made to support me in undertaking my PhD studies and pay the most gratitude especially to my parent who always supports me and encourages me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACC	Accuracy
ANN	Artificial Neural Network
BEEs	Multi-Objective Bees Algorithms
BPSO	Binary Particle Swarm Optimization
CDassign	Crowding Distance
DCT	Discrete Cosine Transform
DEMO	Differential Evolution for Multi-Objective Optimization
DTC-WPT	Dual Tree Complex Wavelet Packet Transform
DTC-WT	Dual Tree Complex Wavelet Transform
ELM	Extreme Learning Machine
ER	Error rate
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
FNDSorting	Fast Non-Dominated Sorting
FR	Feature Ratio
GD	Generational Distance
HV	Hypervolume
IBDFO	Improved Binary Dragonfly Optimization
LPC	Linear Prediction Coefficient
MFCC	Mel-Frequency Cepstral Coefficient
MLP	Multilayer Perceptron
MNSGA-II	Modified Non-Dominated Sorting Genetic Algorithm-II
MODE	Multi-Objective Differential Evolution
MOEAs	Multi-Objective Evolutionary Algorithms
MOP	Multi-Objective Problem
MOPSO- σ	Multi-Objective Particle Swarm Optimization Algorithm using sigma
MOPSO-CD	Multi-Objective Particle Swarm Optimization Algorithm using crowding distance technique
MOPSO-SRD	Multi-Objective Particle Swarm Optimization Algorithm using square root distance
NSGA	Non-Dominated Sorting Genetic Algorithm
NSGA-II	Non-Dominated Sorting Genetic Algorithm-II
PCA	Principle Component Analysis
PNN	Probabilistic Neural Network

RBF	Radial Basis Function
SCA	Sine Cosine Algorithm
SLFN	Single Layer Feed Forward Networks
SP	Spacing
SPEA	Strength Pareto Evolutionary Algorithm
SPEA2	Strength Pareto Evolutionary Algorithm 2
STFT	Short Time Fourier Transform
SVM	Support Vector Machine
VEGA	Vector Evaluated Genetic Algorithm
WHO	World Health Organization
WPT	Wavelet Packet Transform
WT	Wavelet Transform
ZDT	Zitzler-Deb-Thiele
SOP	Single Objective Problem

LIST OF SYMBOLS

Δ	Spread
Ps	Population
n_p	Non-dominated count
PFr_1	Pareto front with number i
P_i	Initial population
Q_i	Updated Population
P_{i+1}	Next iteration population
$Compop$	Combine population
d_i	Euclidean distance
v	Volume

Ciri-ciri Bukan Linear Dan Algoritma Pemilihan Ciri Berdasarkan Berbilang Objektif Untuk Pengelasan Tangisan Bayi

ABSTRAK

Pengiktirafan bayi secara automatik telah menjadi penyelidikan popular dalam dekad yang lalu dan senario ini dapat membantu pengesanan awal status kesihatan bayi. Klasifikasi tangisan bayi mengklasifikasikan status kesihatan bayi berdasarkan isyarat tangisan mereka. Dua pangkalan data telah digunakan, di mana pangkalan data pertama mengandungi 340 sampel Asphyxia, 507 sampel Normal, 879 sampel Deaf, 350 sampel Hungry dan 192 Pain. Pangkalan data lain mengandungi 45 sampel Normal, 531 sampel Premature, dan 513 sampel Jaundice. Ciri-ciri linier seperti Mel Frequency Cepstral Coefficient (MFCC) dan Linear Prediction Coefficient (LPC) biasanya digunakan untuk mengekstrakan maklumat isyarat tangisan yang hanya bergantung pada frekuensi , oleh itu, kaedah pengekstrakan ciri khas masa-frekuensi domain yang dinamakan Dual Tree Complex Wavelet Transform DTC-WT) dan Dual Tree Complex Wavelet Packet Transform (DTC-WPT) digunakan. Sejumlah 2176 ciri khas telah diekstrak daripada keseluruhan metodologi untuk dianalisis. Jumlah ciri khas yang besar mempunyai kecenderungan untuk merumitkan proses perhitungan dan mengurangkan ketepatan dan sering disebut dalam kesusasteraan sebagai laknat dimensi. Untuk mengurangkan dimensi dataset ciri khas yang tinggi, teknik pemilihan ciri khas berasaskan pembungkusan telah dicadangkan untuk memilih ciri khas yang paling relevan. Dua keluaran penting yang terlibat semasa peringkat pemilihan ciri khas adalah bilangan ciri yang dipilih dan ketepatan klasifikasi. Oleh itu, algoritma pengoptimuman pelbagai objektif dicadangkan untuk mengoptimumkan kedua-dua objektif serentak semasa mengurangkan saiz data ciri khas. Kerana tiada crossover dan teknik mutasi tertentu sejagat dalam menyelesaikan semua MOP, juga pertindihan penyelesaian semasa pengoptimuman NSGA-II yang menyebabkan kesulitan mencari penyelesaian optimum. Modified Non-Dominated Sorting Genetic Algorithm-II (MNSGA-II) telah dicadangkan dengan mengadopsi mekanisme pengemaskinian populasi ke dalam NSGA-II konvensional dengan menggantikan teknik crossover dan mutasi dengan fungsi sine kosine dari Sine Cosine Algorithm (SCA). Ujian fungsi Kursawe dan Zitzler-Deb-Thiele (ZDT) digunakan untuk mengesahkan prestasi MNSGA-II yang dicadangkan. MNSGA-II mampu mencapai jarak generasi terendah (GD) 1.22e-03, 1.01e-04, 3.59e-05, dan 1.23e-04 untuk Kursawe, ZDT1, ZDT2, dan ZDT3 yang menunjukkan bahawa penyelesaian yang optimum dijumpai berhampiran dengan Pareto depan yang benar. Akhirnya teknik pemilihan ciri khas secara pembungkusan MNSGA-II telah digunakan untuk menyelesaikan masalah klasifikasi bunyi bayi dalam lima kelas dan tujuh kelas. Satu lagi pangkalan data perubatan (klasifikasi gen kanser microarray) dengan pengkalan data khas yang besar juga digunakan untuk mengesahkan keteguhan MNSGA-II yang dicadangkan. Beberapa binary klasifikasi bunyi bayi telah dijalankan untuk mengkaji prestasi ciri-ciri DTC-WT dan DTC-WPT, ketepatan maksimum 99.85% dan 100% untuk Deaf vs Normal telah dicapai oleh DTC-WT dan DTC-WPT masing-masing. MNSGA-II mengurangkan dimensi ciri khas untuk lima kelas (Hunger, Deaf, Normal, Asphyxia dan Pain) dengan ciri khas sebanyak 511 dan ketepatan pada 96.82%, manakala

573 ciri khas dan ketepatan pada 96.87% untuk pengelasan tujuh kelas (Hunger, Deaf, Normal, Asphyxia, Pain, Jaundice dan Premature). Akhirnya, keputusan dari klasifikasi gen kanser microarray membuktikan keteguhan pemilihan ciri khas berdasarkan pembungkus MNSGA-II dalam corak klasifikasi.

Non-Linear Feature And Multi-Objective Based Feature Selection Algorithms For Infant Cry Classification

ABSTRACT

Automatic infant cry recognition has become popular research in the past decade and this scenario may help in early detection of infant health status. The infant cry classification is classifying the health status of infant based on their cry signals. Two databases were utilized, where first database contains 340 sample of Asphyxia, 507 sample of Normal, 879 samples of Deaf, 350 samples of Hungry and 192 sample of Pain. Another database contains 45 sample of Normal, 531 sample of Premature, and 513 sample of Jaundice. Linear features such as Mel-Frequency Cepstral Coefficient (MFCC) and Linear Predication Coefficient (LPC) are commonly applied in extracting cry signals that only relies on frequency based information, thus, time-frequency domain feature extraction methods named Dual Tree Complex Wavelet Transform (DTC-WT) and Dual Tree Complex Wavelet Packet Transform (DTC-WPT) are employed. A total of 2176 features were extracted from the overall methodology for the analysis. The huge number of features has a tendency to complicate the computation process and reduce accuracy and often cited in literature as curse of dimensionality. In order to reduce the high dimensionality feature dataset, wrapper based feature selection techniques was proposed to select the most relevant features. Two important outputs concerned during feature selection stage are number of selected features and classification accuracy. Therefore, multi-objective optimization algorithms are proposed in optimizing both objectives simultaneously during reducing size of feature dataset. Due to the there is no particular crossover and mutation technique is universally solving all MOP, also duplication of solutions during NSGA-II optimization which cause the difficulty of searching the optimal solutions. A novel Modified Non-Dominated Sorting Genetic Algorithm–II (MNSGA-II) was proposed by adopting the population updating mechanism into conventional NSGA-II by replacing crossover and mutation technique by sine cosine functions from Sine Cosine Algorithm (SCA). Kursawe and Zitzler-Deb-Thiele (ZDT) test functions were apply to validate the performance of proposed MNSGA-II. MNSGA-II able to achieved lowest generational distance (GD) of 1.22e-03, 1.01e-04, 3.59e-05,and 1.23e-04 for Kursawe, ZDT1, ZDT2, and ZDT3 respectively which indicated that the optimal solutions found are near to true Pareto front. Finally MNSGA-II wrapper based feature selection technique was applied to solve five class and seven class infant cry classification problems. Another medical database (microarray cancer gene classification) with huge dimensionality was also used to validate the robustness of proposed MNSGA-II. Several binary infant cry classification were conducted to examine the performance of DTC-WT and DTC-WPT features, maximum accuracy of 99.85% and 100% for Deaf vs Normal were achieved by DTC-WT and DTC-WPT respectively. MNSGA-II reduces the feature dimension for five classes (Hunger, Deaf, Normal, Asphyxia and Pain) with 511 features and accuracy of 96.82%, while 573 features and accuracy of 96.87% for seven classes classification (Hunger, Deaf, Normal, Asphyxia, Pain, Jaundice and Premature). Finally the overall classification results from microarray

cancer gene classification proven the robustness of MNSGA-II wrapper based feature selection in pattern classification.

CHAPTER 1 : INTRODUCTION

1.1 Introduction

Infant cry classification becomes popular research in the past decade. Researchers implemented automation infant cry recognition to differentiate the reason behind infant cries (Lederman, Zmora, Hauschildt, Stellzig-Eisenhauer, & Wermke, 2008). Moreover, the number of the necessary infant cry samples grows exponentially with the feature space dimensionality and huge data storage and computation resources are required. This phenomenon is known as the curse of dimensionality. Feature selection techniques have been introduced to resolve the aforementioned problem. Feature selection has been found to be useful in reducing the computational speed and complexity while improving the accuracy of classification problem (Muthusamy, Polat, & Yaacob, 2015)(Zabidi, Mansor, Lee, Yassin, & Sahak, 2011)(Rosales-Pérez, Reyes-García, Gonzalez, & Arch-Tirado, 2012)(Saeys, Inza, & Larrañaga, 2007)(Yu & Liu, 2004). Obtaining robust features in biomedical and biometric application is a difficult task due to features being influenced by inherent factors can cause further difficulties in feature extraction.

Metaheuristic optimization deals with optimization problems using metaheuristic algorithms. Optimization is basically everywhere, from engineering design to medical application. As time and resources are always the main concern in real-world optimization, hence the optimal utility of these criteria are crucially important. Most of the optimization objectives are conflict and incommensurable with each other, it is not an easy task to search for the optimal solutions. Benchmarks studies and biomedical problems were researched and an attempt was made to improve the classification accuracy of these experiments by introducing new feature extraction and selection model after considering the shortcomings of the existing approaches. Two widely used multi-objective benchmarks (Kursawe test function and Zitzler-Deb-Thiele (ZDT) test functions) were tested with the modified multi-objective algorithms to ensure the performance ability. Then, we focus on infant cry classification and existing medical datasets for biomedical problems by implementing the proposed multi-objective algorithm to reduce the feature size while improving the classification accuracy.

1.2 Problem Statement and Motivations

According to statistics 2015, 1 million neonatal deaths occur on the day of birth and close to 2 million dies in the first week of life (WHO, UNICEF, UNFPA, Group, & UNPD, 2015). Every year, 15 million babies are born before 37 completed weeks of gestation (preterm) and this number is rising. The rate of preterm birth ranges from 5% to 18% of babies born across 184 countries. ("Preterm birth," 2018). All the pathological infant could have been treated or given treatment and therapies if diseases would have been detected earlier. The research of cry signal is to look for differences of atypical cry signals of infants who are at risk of several diseases and developmental disorders. In previous studies, acoustic features of infant cry are reported to change according to infant's physiological state (Fuller, Keefe, Curtin, & Garvin, 1994; Goberman & Robb, 2005; Green, Gustafson, & McGhie, 1998). Hence, newborn infant cry signals are believed to be useful to assist clinical diagnosis.

Many studies have been carried in recent year to effectively classify the different types of cry signals. Mel Frequency Cepstral Coefficient (MFCC) (Chittora

& Patil, 2014; Farsaie Alaie, Abou-Abbas, & Tadj, 2016; Zabidi, Lee, Mansor, Yassin, & Sahak, 2010) and Linear Prediction Coefficient (LPC) (Galaviz & García, 2005; Orozco García & Reyes-García, 2003a; Rosales-Pérez et al., 2015) are commonly used in infant cry classification. However, MFCC and LPC are limited in analysing the non-stationary characteristics of infant cry signals (Saraswathy et al., 2018). In application involving infant cry signals, researchers found that time-frequency based techniques are performing better compared to time or frequency domains methods. Briefly, it can be said that, in the literature, different interpretations and representations of infant cry signals using some time-frequency based techniques namely wavelet packet transform (WPT), and short time Fourier transform (STFT) and the outcomes from the previous works persuaded to explore and investigate more on the effectiveness of time-frequency methods in infant cry classification (Hariharan, Saraswathy, Sindhu, Khairunizam, & Yaacob, 2012; Hariharan, Sindhu, & Yaacob, 2012; Saraswathy, Hariharan, Vijean, Yaacob, & Khairunizam, 2012).

In medical studies, the development of computer and data collection technologies have increase the variables or features of the datasets in pattern recognition. Conventionally, to process such compact datasets is tedious and time consuming for conventional machine learning. The datasets contained some features or variables that are irrelevant or redundant to the signal that may cause the misclassification during the classification process. Feature selection techniques have been broadly proposed in many cutting-edge research areas to reduce the number of features in order to increase the classification accuracy. Two main objectives have to optimize during feature selection which are the number of selected features and classification accuracy. The main difference between the single-objective optimization problem (SOP) and the multi-objective problem (MOP) is that the solution of SOP is usually a single optimum solution corresponding to the objective function, but the MOP aims to optimize more than one objective function, thus yielding a set of solutions which are non-dominated to one another in context of their corresponding objective values. In (Hariharan et al., 2018), a wrapper based feature selection in infant cry classification with both the objectives (number of selected features and classification accuracy) are converted into single objective by using weighted sum method. However, the weighted sum method has a drawback: inadequate choice of weight factors may cause the incommensurable objectives to lose their significance on combining into a single objective function (Jeyadevi, Baskar, Babulal, & Willjuice Iruthayarajan, 2011).

ZDT and Kursawe test function are two well-known benchmark studies used to evaluate multi-objective evolutionary algorithms. So far, there is many research conducted to improve the search of Pareto optimality using these test functions as benchmarks testing. For instance, Coello (2001) has proposed a micro evolutionary algorithm to solve Kursawe test function (Coello et al., 2007). Due to the intrinsic mathematical properties, these functions are employed to reveal the challenging issues of MOPs. According to Zitzer et al. (2000), ZDT test function can cause difficulty of Pareto convergence whereas Kursawe test function has a non-connected Pareto front. Considering these challenges, both ZDT and Kursawe functions are used as the benchmark evaluations for our research validation.

There are few research gaps found which are the feature extraction methods, feature selection techniques and multi-objective optimization algorithms. Firstly, feature extraction techniques are widely applied in infant cry classification using time domain analysis such as MFCC and LPC, but these methods only focused on time domain, thus time frequency based techniques should be explored and investigate more on the effectiveness of time-frequency methods in infant cry classification.

Next, there are two main objectives need to be optimized during feature selection process named number of feature and classification accuracy, these objectives are combined to form single objective and weighted sum method is applied in (Hariharan et al., 2018). Due to the inadequate choice of weight factors might cause these objectives lose their significance, the multi-objective methods are considered to optimized the objectives without biasing to any objective during optimization.

Finally, there is no individual evolutionary algorithm claimed to be the best algorithm and solve all the MOPs, therefore, always an improved/hybrid version of multi-objective evolutionary algorithms are designed (S. M. Mirjalili, Saremi, Mirjalili, & Coelho, 2016; Toscano Pulido & Coello Coello, 2003) By considering the combination of advantages from different algorithms, there are motivations for the development of improving algorithm by combining several learning strategies for improvements in solving multi-objective problems. As a result, this study explores the elitism of available evolutionary algorithms and invents a new optimization algorithm to enhance the Pareto optimality coverage for MOPs (Jeyadevi et al., 2011; Ramesh, Kannan, & Baskar, 2012)

Based on the abovementioned challenges and motivations, we develop an improved evolutionary algorithm, MNSGA-II, by integrating a number of meta-heuristic strategies from the previous research considering the strength of different algorithms. Next, our proposed optimization model then applied on ZDT and Kursawe benchmark functions to investigate the capability to obtain better Pareto optimality. Finally, the model is used to develop a multi-objective wrapper-based feature selection technique to reduce the feature dimensionality and assist infant pathological classification to identify the pathology symptoms.

1.3 Objectives of this Study

This research aims to achieve the following objectives:

i. To develop a non-linear feature extraction method for the extraction of infant cry signal.

ii. To develop a new multi-objective optimization algorithm

iii. To validate the performance of proposed multi-objective optimization algorithm in benchmark test function studies and as wrapper-based feature selection techniques in pattern recognition applications.

1.4 Project Scope

The main scope of the research work is to investigate the performance of non-linear features and develop a modified multi-objective wrapper based feature selection in infant cry classification. This study performs an investigation of the properties and advantages of Multi-Objective Evolutionary Algorithms (MOEAs) to develop a novel hybrid evolutionary optimization model. The benchmark studies considered in this research are concerned with the capability to find Pareto optimality. Thus, the research focuses on ZDT and Kursawe test functions which are challenging for finding good Pareto optimality. Based on the findings, a multi-objective wrapper-based feature selection technique is developed to assist pattern recognition. Some of the medical datasets selected consists of thousands of variables or features and this is a challenging task for feature selection algorithm to identify the relevant feature subsets.

1.5 Thesis Outline

This thesis is composed of five chapters, the content is organized as follows:

Chapter 1 discusses the non-linear features and wrapper based feature selection in infant cry classification, the problem statements, the research objectives and project scope.

Chapter 2 elaborates the comprehension of pattern recognition, the Multi-Objective Problems, and Multi-Objective Evolutionary Algorithm. Also the review on Multi-Objective test functions.

Chapter 3 describes the proposed methodology in this research, including the non-linear feature extraction techniques, the Modified Non-Dominated Sorting Genetic Algorithm, the benchmark test functions and biomedical datasets.

Chapter 4 presents the results of the proposed non-linear features in infant cry classification, the validation of proposed modified algorithm, multi-class infant cry classification results, and microarray cancer gene expression classification.

Chapter 5 concluded the findings of this study, also the limitation and future works.