FEATURE SELECTION METHOD BASED ON HYBRID FILTER-METAHEURISTIC WRAPPER APPROACH

NEESHA A/P JOTHI

UNIVERSITI SAINS MALAYSIA

2020

FEATURE SELECTION METHOD BASED ON HYBRID FILTER-METAHEURISTIC WRAPPER APPROACH

by

NEESHA A/P JOTHI

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosphy

November 2020

ACKNOWLEDGEMENT

First and foremost, I would like to express my greatest gratitude to my supervisors Dr.Sharifah Mashita, PM Wahidah Husain, and Dr.Nur'Aini Abdul Rashid from the School of Computer Sciences of Universiti Sains Malaysia for their consistent supervision provided throughout this study, invaluable guidance, encouragement, positive comments and continuous assistance and support they have rendered throughout these years in completing my thesis. Both PM Wahidah and Dr.Nur'Aini have been strong support since the beginning of my studies and they made their presence strong even though they have retired. Special thanks to Dr.Sharifah Mashita for helping me to gather my missing motivations, and to put everything through to make the thesis completion. She has indeed helped me a lot towards the later stage of my studies.

I would also like to express great reverence for my parents Jothi Rengasamy and Jayakumari Kali for being a great pillar of strength since the day I embarked on this journey, they have made everything possible for me. My studies years wouldn't have been possible without their immense support and encouragement. Special thanks to my siblings for always being there and helpful whenever I needed them. I would like to say a special thanks to Cheenu for always encouraging, listening to me and being there always whenever I needed someone to talk too.

I would like to thank the Universiti Sains Malaysia for providing the RUI grant of "Enhanced Nature Inspired Based Classification Algorithm for Medical Data Analysis" with grant number 1001/PKOMP/8014076, and MyBrain15 scholarship respectively. I gratefully acknowledge the assistance of administrative assistant from the School of Computer Sciences, Sheela Muniandy and many others whom the names I

did not mention here. Finally, I would like to thank my friends near and dear for being a great support throughout my duration of studies as all your kind acts and gestures meant immensely to me.

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION

CHAPTER 2 LITERATURE REVIEW

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

KDD	Knowledge Discovery in Database
MRMR	Minimum Redundancy Maximum Redundancy
SFS	Sequential Forward Strategy
SBS	Sequential Backward Strategy
SVM	Support Vector Machine
DA	Dragonfly Algorithm
SA	Simulated Annealing
PCA	Principal Component Analysis
LDA	Linear Discriminant Analysis
mRMR	Minimal Redundancy Maximal Relevance
SU	Symmetric Uncertainty
GA	Genetic Algorithm
NB	Naïve Bayes
FHR	Fetal Heart Rate
AIC	Akaike Information Criteria
BIC	Bayesian Information Criterion
PSO	Particle Swarm Optimization
RBF	Radial Basis Function
WNN	Wavelet Neural Network

AMI	Acute Myocardial Infarction
ELM	Extreme Machine Learning
MI	Mutual Information
CS	Cosine Similarity
PCC	Pearson correlation coefficient
JS	Jaccard similarity
KNN	K-Nearest Neighbor
ABC	Artificial Bee Colony
GBC	Genetic Bee Colony
CRO	Coral Reef Optimization

KAEDAH PEMILIHAN CIRI BERDASARKAN PENDEKATAN GABUNGAN PENAPIS-PEMBALUT METAHEURISTIK

ABSTRAK

Data berdimensi tinggi sering dikaitkan dengan ciri yang tidak diperlukan dan terdapat banyak pendekatan maklumat berteori yang digunakan untuk memilih perkumpulan ciri yang paling relevan dan untuk mengurangkan saiz ciri-ciri tersebut. Tiga pendekatan yang paling bererti adalah pendekatan penapis, pendekatan pembalut, atau pendekatan terbenam. Kebanyakan pendekatan penapis gagal untuk mengenal pasti sumbangan individu bagi setiap ciri dalam setiap perkumpulan ciri dalam mencapai subset ciri yang optimum. Sementara itu pendekatan pembalut mempunyai masalah pada interaksi yang kompleks di antara ciri-ciri dan genangan dalam optima tempatan. Untuk menangani, kelemahan ini, kajian ini menyiasat pendekatan penapis dan pembalut untuk membangunkan pendekatan hibrid yang berkesan untuk pemilihan ciri. Kajian ini tertumpu kepada kaedah yang mempunyai dua peringkat untuk memilih subset ciri yang paling optimum dalam masalah pemilihan ciri. Tahap pertama dalam kajian ini telah mencadangkan kaedah hibrid berdasarkan Nilai ReliefF-Shapley sebagai pemilihan ciri untuk mengenal pasti sumbangan ciri individu dalam mencapai subset ciri yang optimum sebagai sumbangan pertama tesis ini. Walau bagaimanapun, semasa pencarian di peringkat ini, kaedah yang telah dicadangkan menghadapi beberapa permasalahan dalam pemilihan subset ciri yang optimum disebabkan oleh sifat asal algoritma yang berfungsi sebagai algoritma carian global. Terdapat banyak pendekatan berasaskan metaheuristik yang diketahui telah digunakan sebagai kaedah pencarian untuk carian global dalam pendekatan pembalut. Walau bagaimanapun, kebanyakan pendekatan ini telah mengalami genangan dalam optima tempatan yang disebabkan oleh interaksi yang rumit antara ciri-ciri dan ruang carian yang besar. Oleh itu, sumbangan kedua tesis ini adalah tertumpu kepada hibridisasi Algoritma Pepatung (DA) dan Simulasi Penyepuhlindapan (SA) sebagai algoritma carian tempatan untuk meningkatkan eksploitasi tempatan bersama dengan sumbangan ciri individu dari peringkat pertama. Kaedah yang telah dicadangkan dibandingkan dengan kaedah-kaedah lain daripada penelitian literatur dengan menggunakan 11 dataset tanda aras dengan saiz dan dimensi yang berbeza. Ketepatan pengelasan yang diperoleh daripada Nilai ReliefF-Shapley adalah melebihi 75% berbanding empat kaedah terkini yang telah dinilai. Kaedah yang dicadangkan telah mendapat ketepatan pengelasan sebanyak 81.56% menggunakan Musk, 83.60% menggunakan Optical, 79.11% menggunakan Arrhythmia, 81.98% menggunakan Isolet, 87.81% menggunakan Arcene, 97.97% menggunakan Dexter, 82.12% menggunakan Dorothea, 84.77% menggunakan Gisette, 83.68% menggunakan Glass, 74.44% menggunakan Kanser paru-paru, dan 90.86% menggunakan Madelon. Manakala, ketepatan pengelasan yang telah diperolehi daripada DA-SA dengan Nilai ReliefF-Shapley berada di atas 80% berbanding dengan tiga kaedah terkini yang telah dinilai. Kaedah yang dicadangkan telah mendapat ketepatan pengelasan sebanyak 96.30% menggunakan Arcene, 92.67% menggunakan Arrhythmia, 83.12% menggunakan Dexter, 84.39% menggunakan Dorothea, 87.70% menggunakan Gisette, 91% menggunakan Glass, 59.63% menggunakan Isolet, 94% menggunakan Kanser paru-paru, 71% menggunakan Madelon, 65% menggunakan Musk dan 85% menggunakan Optical. Kaedah yang telah dicadangkan dapat menentukan keputusan persaingan pada kebanyakan dataset dan telah mencapai hasil yang terbaik pada beberapa dataset serta meningkatkan prestasi klasifikasi pengelasan.

FEATURE SELECTION METHOD BASED ON HYBRID FILTER-METAHEURISTIC WRAPPER APPROACH

ABSTRACT

High dimension data are often associated with redundant features and there exist many information-theoretic approaches used to select the most relevant set of features and to reduce the feature size. The three most significant approaches are filter, wrapper, and embedded approaches. Most filter approaches fail to identify the individual contribution of every feature in each set of features in achieving an optimal feature subset. While the wrapper approaches encounter problems from complex interactions among features and stagnation in local optima. To address, these drawbacks, this study investigates filter and wrapper approaches to develop effective hybrid approaches for feature selection. This study focuses on a two-stage method to select the most optimal subset of features in the feature selection problems. The first stage of this study proposed a hybrid method based on ReliefF-Shapley Value as feature selection to identify the individual feature contribution in achieving an optimal feature subset as the first contribution of this thesis. However, during the searching in this stage, the proposed method faced some issues in the selection of the optimal feature subset due to the nature of the algorithm in performing as a global search algorithm. There are many metaheuristic-based approaches have known to be implemented as a searching method for global search in the wrapper approach. However, most of these approaches experience stagnation in local optima which is caused by complex interactions among features and large search space. Therefore, the second contribution of this thesis focuses on hybridizing the Dragonfly Algorithm (DA) and Simulated Annealing (SA) as a local search algorithm to enhance local exploitation along with individual feature contribution from the first stage. The proposed methods are compared with other methods from literature using 11 benchmark datasets with different sizes and dimensions. The classification accuracy obtained from ReliefF-Shapley Value was above 75% against the four state-of-the-art methods evaluated. The proposed method obtained a classification accuracy of 81.56% on Musk, 83.60% on Optical, 79.11% on Arrhythmia, 81.98% on Isolet, 81.56% on Musk, 87.81% on Arcene, 97.91% on Dexter, 82.12% on Dorothea, 84.77% on Gisette, 83.68% on Glass, 74.44% on Lung cancer, and 90.86% on Madelon. Whereas the classification accuracy obtained from DA-SA with ReliefF-Shapley Value was above 80% against the three state-of-the-art methods evaluated. The proposed method obtained a classification accuracy of 96.30% on Arcene, 92.67% on Arrhythmia, 83.12% on Dexter, 84.39% on Dorothea, 87.70% on Gisette, 91% on Glass, 59.63% on Isolet, 94% on Lung cancer, 71% on Madelon, 65% on Musk, and 85% on Optical. The proposed methods ascertain competitive results on most of the datasets and achieved the best results on some datasets as well as improving the classification performances.

CHAPTER 1

INTRODUCTION

1.1 Background

Knowledge Discovery in Database (KDD) is the overall process of discovering useful knowledge, in which the process has continued to evolve from the intersection of various research domains, including artificial intelligence, databases, machine learning, statistics, data visualisation, pattern recognition, high-performance computation, and knowledge acquisition for expert systems (Khosrow-Pour, 2019; Fayyad et al., 1996). The KDD applies the database with the required selection comprising of pre-processing, sub-sampling, and transformation to apply data mining techniques to enumerate patterns. Based on the data and discerned patterns, the products of data mining were assessed in this study to identify the subset of enumerated patterns viewed as knowledge (Khosrow-Pour, 2019; Fayyad et al., 1996).

The KDD denotes the overall process of discovering useful knowledge from data, whereas data mining is a step in this process. Data mining is the application of certain algorithms to extract patterns from data within the KDD process. The KDD process comprises of data preparation, data selection, data cleaning, incorporation of suitable prior knowledge, and proper interpretation of the mining outcomes, which are essential to ascertain retrieval of useful knowledge from the data. The KDD process is illustrated in Figure 1.1. Data mining refers to one of the most rapid uprising subject matter within the information domain stemming from the vast data available and the need to translate the data into useful information (Khosrow-Pour, 2019; Agarwal, 2013; Han et al., 2012). Being integral in KDD process, data mining is comprised of an iterative sequence of tasks for data pre-processing and mining, follows by pattern evaluation and knowledge presentation (Khosrow-Pour, 2019; Agarwal, 2013; Han et al., 2012). The pre-processing step in data mining is performed prior to the same step in KDD process due to its huge effect on the performances. Another important step in pre-processing for data mining is feature selection. The feature selection is often used to discards irrelevant and repeating variable in the dataset (Zawbaa et al., 2018; Huan & Motoda, 1999). Irrelevant features do not provide any valuable information, while those redundant do not offer any additional information (Zawbaa et al., 2018; Kalousis et al., 2006). The final step in the KDD model is interpretation/evaluation. This step is important in finding interesting patterns, summarizing, and visualizing them to make the data more understandable by users.

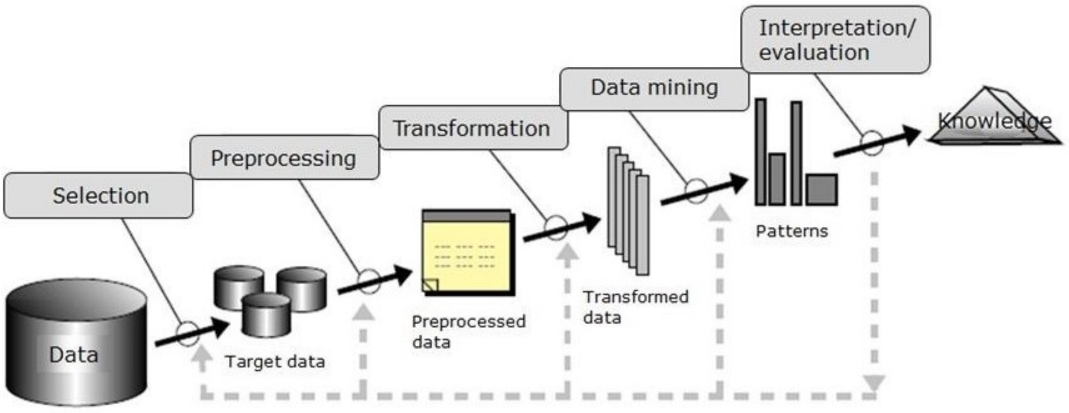

Figure 1.1: The KDD Process (Fayyad et al., 1996)

From the literature, there have been three approaches to feature selection are wrapper, hybrid, and filter approaches (Solorio-Fernández et al., 2019; John et al., 1994).

The hybrid approach in feature selection methods integrates filter and wrapper approaches. Such integration aids in detecting informative features that have high accuracy for classification (Solorio-Fernández et al., 2019; Guyon & Elisseeff, 2003.

In the filter approach, the features are applied for data training to assess the importance of the features or the feature subset (Solorio-Fernández et al., 2019; Kohavi & John, 1997). Machine learning algorithms are excluded in this approach to eliminate features that are redundant and irrelevant (Solorio-Fernández et al., 2019; John et al., 1994). Each feature in the filter approach depends on varying metrics of multiple attributes, including information theory, probability, distribution, and distance. Various subsets of features are generated from the dataset found in every filter. This results in varied performances after employing machine learning algorithms (Seijo-pardo, 2016; Bolón-Canedo et al., 2014). Exceptional outcomes generated by the filter approach may vary from those of another dataset, thus giving highly variable classification performance results. Put simply, estimation and predictiveness are lacking in the chosen subset of features. Some of the common techniques used in filter approach are Chi-Square (Thaseen et al., 2018; Maben & Sharp, 2001), Fisher Score (Saqlain et al., 2018; Tsuda et al., 2002), Information Gain (C.-M. Lai et al., 2016; Hu et al., 2013), Minimum Redundancy Maximum Redundancy (MRMR) (X. Yan & Jia, 2019; Peng et al., 2005) and ReliefF (Angadi & Reddy, 2019; Robnik-Šikonja & Kononenko, 2003).

The wrapper approach works by evaluating a subset of features using a machine learning algorithm that employs a search strategy to look through the space of possible feature subsets, evaluating each subset based on the quality of the performance of a given algorithm (Jantawan & Tsai, 2014; A. Jain & Zongker, 1997). The two stages involved to classify feature subset are searching and evaluation. Search-based approaches are performed at the search stage to produce a subset with discriminative features in accordance with effective classification (Jantawan & Tsai, 2014; A. Jain & Zongker, 1997). Finding an optimal subset of features is also called a nondeterministic polynomial (NP) hard problem (Jantawan & Tsai, 2014; A. Jain & Zongker, 1997). Approaches based on metaheuristic techniques are applied to function as a search method in the wrapper approach (Alshamlan et al., 2015). The wrapper approach is comprised of two main components: a) subset generation (performs search methods), and b) evaluation (performs machine learning algorithm). Some search methods employed for subset generation are Sequential Forward Strategy (SFS) (Marcano-Cedeno et al., 2010), Sequential Backward Strategy (SBS) (Olvera-López et al., 2010), and Genetic Algorithm (GA) (Sayed et al., 2019; Olvera-López et al., 2010). The evaluation process embeds machine learning algorithms, such as Support Vector Machine (SVM) (Tuba et al., 2019; Cortes & Vapnik, 1995), Naïve Bayes (NB) (Bashir et al., 2019; Kelemen et al., 2003) and K-Nearest Neighbour (KNN) (Atallah et al., 2019; Marchiori, 2013). The machine learning algorithms are employed in the wrapper approach to determine feature subsets.

Some approaches have hybridised the filter and wrapper approaches (Solorio-Fernández et al., 2019;Guyon & Elisseeff, 2003). Metaheuristic approaches have been used extensively as a hybrid approach, along with the wrapper approach alone, to overcome issues in the feature selection. The performance of the metaheuristic algorithms has been proved to be one of the best performing techniques, which have been extensively used for solving feature selection problem (Salem et al., 2017; Mahajan et al., 2016; Aydilek & Arslan, 2013). Works of the literature reveal some metaheuristic-based methods used to address issues related to feature selection, including Correlation-based Feature Selection with improved Binary Particle Swarm Optimisation (iBPSO) and a Combat GA known as (IG-SGA) (Salem et al., 2017), Genetic Algorithm (GA) with Artificial Bee Colony (ABC) (Alshamlan et al., 2015) and Hybrid Whale Optimisation Algorithm (WOA) with SA for Feature Selection (M. M. Mafarja & Mirjalili, 2017). Nonetheless, most techniques face stagnation within local optima due to intricate interaction between massive search space and features (I. Jain et al., 2018). The metaheuristic approach, which refers to a process for iterative enhancement, employs operators and pools all knowledge related to a certain problem to assess the search space and attain solutions that are viable (Osman & Laporte, 1996). In metaheuristic approaches, search space refers to a bounded domain that embeds all viable solutions to address a particular problem.

The metaheuristic approach should find a balance between exploration and exploitation while the search process. During exploration, unvisited new areas in the search space are explored, whereas exploitation performs an intensive search in already-visited areas. The two groups of metaheuristic approaches are local search and population-based techniques. The first technique, local search, is a solution at a time that offers enhancement by employing the structures of the neighbourhood. Its primary benefit is the search speed, while the drawback is that it can get stuck easily in local optima due to the emphasis on exploitation. Instances of local search-based technique are Simulated Annealing (SA) (Fathollahi-Fard et al., 2019; Kirkpatrick et al., 1983), and Tabu Search (Sivaram et al., 2019; Glover, 1989). In the population-based technique, the solution population is weighed in at a time to re-combine existing solutions in creating a new solution(s) at iteration with emphasis on exploration. The examples of this technique are GA (Salem et al., 2017; Harik et al., 1999), Ant Colony Optimisation (L. Sun et al., 2019; Blum, 2005), and Harmony Search (HS) Algorithm (Sudha & Selvarajan, 2019; Yang, 2009).

1.2 Motivations

The high dimensional data can contain a high degree of irrelevant and redundant information which may greatly degrade the performance of learning algorithms. Therefore, the feature selection becomes very necessary for machine learning tasks when facing high dimensional data nowadays (Zawbaa et al., 2018; Yu & Liu, 2003). However, this trend of enormity on both size and dimensionality also poses severe challenges to feature selection algorithms (Zawbaa et al., 2018; Yu & Liu, 2003). These traits pose a challenge to the data interpretation and analysis and for computational learning algorithms to produce accurate accuracy. From a computational point of view, finding informative features and ignoring irrelevant and redundant features are challenging tasks. However, the feature selection help to enhance the predictive accuracy of classifier techniques and are able to interpret the pattern of the selected features (Solorio-Fernández et al., 2019; Dash & Liu, 1997). Nevertheless, the existence of a large number of features is a challenging issue in the development of an efficient classifier called machine learning algorithm (C.-M. Lai et al., 2016). To address this challenge and to improve the predictive accuracy, this study applies a feature selection approach which is a data pre-processing step in data mining to find the subset of a most optimal subset of features which can provide enhanced classification accuracy (Zawbaa et al., 2018; A. Jain & Zongker, 1997).

1.3 Problem Statement

Feature selection is one of the approaches used to achieve dimensionality reduction of data to improve machine learning performance. Feature selection algorithms are also aimed at finding the optimal subset that will optimise the performance of machine learning algorithms. According to Huan Liu (2004), a feature can be classified into three categories: 1) strongly relevant, 2) weakly irrelevant, and 3) irrelevant. A strongly relevant feature designates that the feature is necessary for achieving an optimal subset. The strongly relevant features cannot be removed as they affect the original condition of the class distribution (Huan Liu, 2004). The weakly irrelevant features are not always necessary to achieve an optimal subset at certain conditions (Huan Liu, 2004). The irrelevant features are not necessary at all in achieving an optimal subset (Huan Liu, 2004). The feature selection methods should have an optimal subset of features that should include all the strongly relevant features, a subset of weakly relevant features, and none of the irrelevant feature as result (Huan Liu, 2004). There is no proper definition on which weakly relevant features should be included and which of them should be excluded (Zawbaa et al., 2018). Therefore, it is necessary to have a clear definition of the said three categories before performing the feature selection techniques. The filter approach depends on varying metrics of multiple attributes, including information theory, probability, distribution, and distance. Various subsets of features are generated from the dataset found in every filter. This results in varied performances after employing machine learning algorithms (Seijo-pardo, 2016; Bolón-Canedo et al., 2014). Exceptional outcomes generated by the filter approach may vary from those of another dataset, thus giving highly variable classification performance results. However, the conventional feature selection techniques based on the filter approach tend to dismiss vital interdependent structure of features during the selection of optimal feature subset (Xin et al., 2013; Xin Sun et al., 2012). The illustration of the problem statement is shown in Figure 1.2. This happens because most of the subset evaluation processes, which are based on information measurements, have failed to identify how the features collaborate and their contribution towards the optimal subset of features.

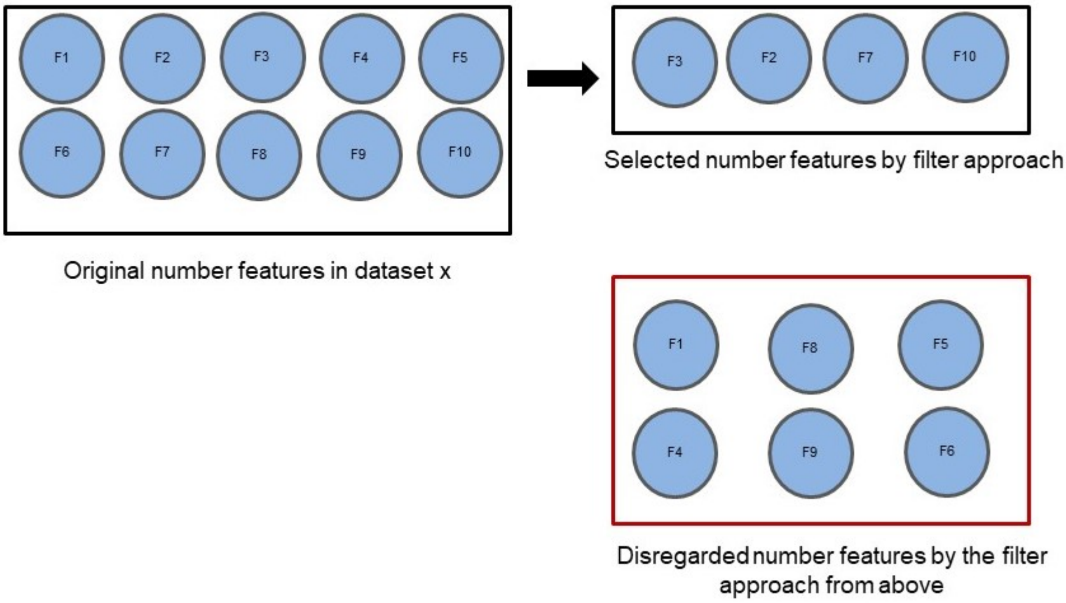
F1
F2
F3
F4
F5
F6
F7
F8
F9
F10
Original number features in dataset x
F3
F2
F7
F10
Selected number features by filter approach
F1
F8
F5
F4
F9
F6
Disregarded number features by the filter approach from above

Figure 1.2: The illustration of the first problem statement

Meanwhile, in the wrapper approaches, the classification of the feature subset is achieved through two ways: a) searching and b) evaluation. During the searching stage, search-based methods are utilized to generate a discrimination feature subset based on an efficient classification ((Jantawan & Tsai, 2014); (A. Jain & Zongker, 1997)). Having said that, the metaheuristic-based approaches have known to be implemented as a searching method in a wrapper approach including the population-based algorithms (Alshamlan et al., 2015). However, most of these approaches experience stagnation in local optima caused by complex interactions among features and large feature search space during their searching process (I. Jain et al., 2018). The illustration of the said problem is shown in Figure 1.3

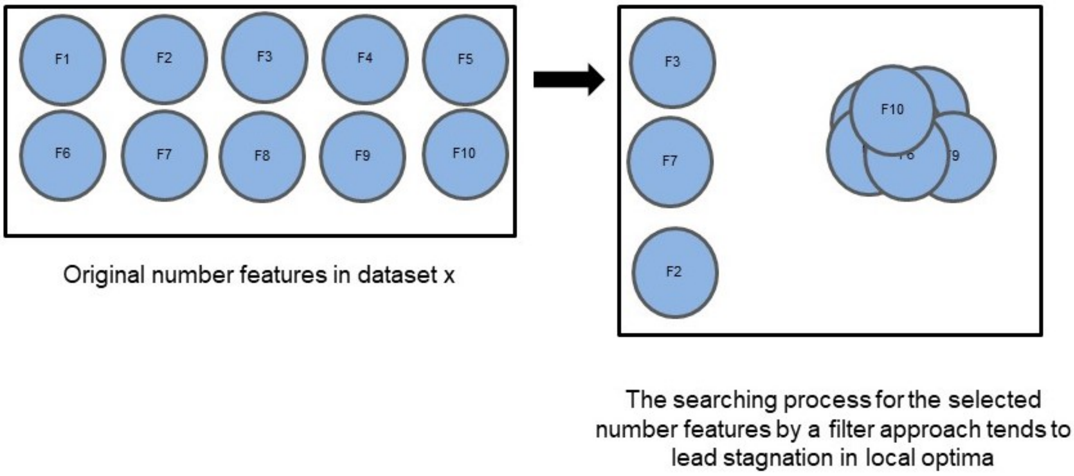
F1
F2
F3
F4
F5
F6
F7
F8
F9
F10
Original number features in dataset x
F3
F7
F2
F10
9
The searching process for the selected number features by a filter approach tends to lead stagnation in local optima

Figure 1.3: The illustration of the second problem statement

1.4 Research Objectives

To achieve the research goal, the objectives of this research are:

1. To propose a hybrid feature selection method based on filter and wrapper approaches in selecting the most optimal subset of features in the feature selection problem.

2. To enhance the performance of filter and wrapper based approach by proposing a hybrid of local search and population-based algorithms to improve local exploitation.

1.5 Research Contributions

Some contributions of this study are detailed in the following:

1. A hybrid of ReliefF and Shapley Value is proposed to select the most essential features for feature selection.
2. A hybrid of Dragonfly Algorithm (DA) with Simulated Annealing (SA) is prescribed to provide a more structured local search, and hence, achieve maximum exploitation.

1.6 Thesis Outline

The rest of the report is organised into six chapters, as described in the following: Chapter 2 presents a review of the related literature relevant to this study. Chapter 3 depicts the research methodology that illustrates every stage taken in this study. Chapter 4 of the thesis discusses the first contribution which is the hybrid filter-wrapper method known as ReliefF-Shapley Value developed to select the optimal subset of features. Chapter 5 of the thesis prescribes the second contribution which is the hybrid method known as DA-SA developed to attain a solution to address the local optima issue faced while selecting the optimal subset of features. The final chapter concludes the study and provides some recommendations for future undertakings.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter introduces the literature review used to lay down the fundamental ideas which lead to accomplishing the research objectives. The structure of the literature review that is covered in this thesis in Figure 2.1. This chapter is organised into various sections which consist of the overview of feature selection, the feature selection processes, feature selection approaches, findings from the literature review, and a research gap. The overview of feature selections provides an overview of feature selection. The feature selection processes section discusses the generic feature selection process. The feature selection approaches discuss the various work used for the review purpose. The findings from the literature review present the summary of the papers in a tabular format and a research gap. The chapter summary is presented in the final section.

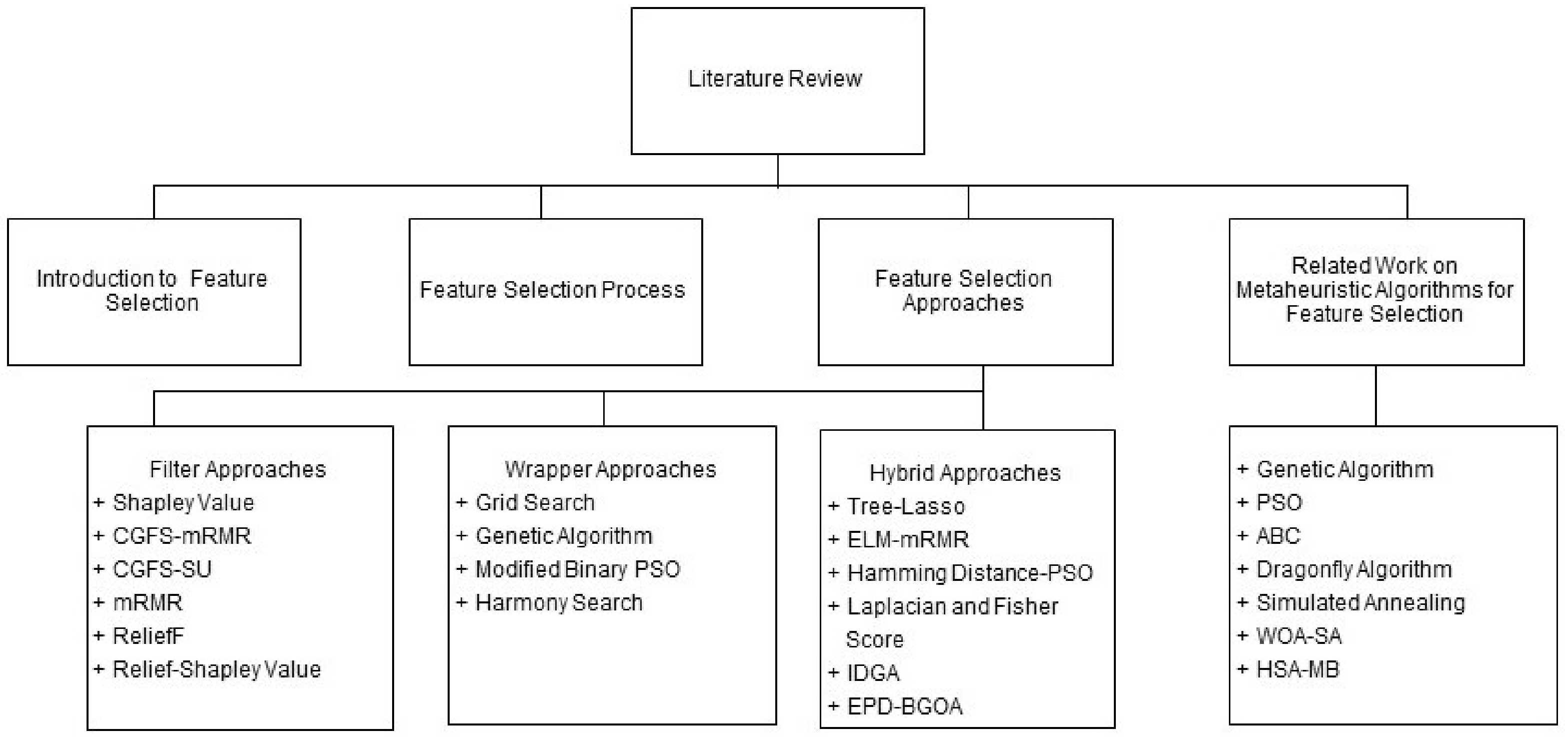

Figure 2.1: Structure of the literature review

2.2 An overview of Feature Selection

Feature selection from datasets is performed due to two reasons. First, it is to minimise the dataset size for effective analysis. Second is for dataset adaptation in selecting the most suitable analysis approach (Kumar & Minz, 2014). The first reason is of utmost importance due to the vastly available techniques to date. Besides, the dataset size tends to grow larger in dimension and volume. Reduction of dataset size is possible via sample set and feature set reduction. The high amount of dataset features, which is relative to or exceeding the samples, can lead to overfitting of the model, thus generating poor outcomes for the validation dataset. Besides, high computation needs must be satisfied to develop a classification model from a dataset that consists of multiple features ((I. Jain et al., 2018); (Kumar & Minz, 2014)). The reduction may be carried out via feature selection and extraction (transformation). Some techniques of feature extraction are Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA), and Multidimensional Scaling methods to transform existing features into a new set in accordance to their combinations, in order to detect meaningful data (Jovic et al., 2015). This new set of features may be decreased easily by weighing in some characteristics, including the convergence of data variance.

Meanwhile reduction for feature set is dictated by feature redundancy and relevance. The three categories of features are strongly relevant, weakly irrelevant, and irrelevant (Gore & Govindaraju, 2016; Kumar & Minz, 2014; Xin Sun et al., 2012; Huan Liu, 2004). The strongly relevant feature is significant for an optimal subset of features, which if eliminated can affect the distribution of the original conditional target (Xin Sun et al., 2012; Huan Liu, 2004). Meanwhile, a weakly irrelevant feature may be indifferent to an optimal subset but relying on some conditions (Gore & Govindaraju, 2016; Huan Liu, 2004). Lastly, irrelevant features may be dismissed altogether (Huan Liu, 2004). Identifying redundancy in multivariate cases is essential upon assessing feature subset (Molina et al., 2002). Minimising redundancy and maximising relevance are the goals of feature selection. Relevant features are sought in seeking feature subset (Huan Liu, 2004). In meeting the goal of optimal feature subset, the technique of feature selection can examine in total, $2m - 1$ subsets, where m refers to the total number of features in the dataset with the exclusion of empty subset (A. Jain & Zongker, 1997). This is not feasible in terms of computation even for a moderately huge dataset. This led to the application of heuristic approaches in seeking viable subsets, thus putting the completeness of the search aside. There are four steps in the feature selection process. The four steps are: a) subset generation, b)subset evaluation, c)stopping criteria, and d)validation, they are elaborated in the next section.

2.3 Feature Selection Process

The feature selection process can be categorised into four major steps: a) subset generation, b) subset evaluation, c) stopping criterion, and d) result validation. Subset generation is also known as a search strategy that produces feature subsets for evaluation based on a certain search strategy. There are three generic strategies which are the complete search, heuristic search, and random search which are discussed in detail in sub-section 2.3.1(a) - 2.3.1(c). Then, each of the generated subsets is evaluated and compared with the previous best one accordingly to a certain evaluation criterion. There are five categories of the subset evaluation : distance, information, dependence, consistency, and classifier error rate, which are discussed in detail in sub-section 2.3.2(a) - 2.3.2(e). If the new subset turns out to be better, it replaces the previous best subset. The process of subset generation and evaluation is repeated until a given stopping criterion is satisfied. Then, the selected best subset is validated by prior knowledge or different test via synthetic and/or real-world data set (Huan Liu, 2004; Guyon & Elisseeff, 2003). The general process of feature selection is shown in Figure 2.2. This is a generic process of the feature selection process, the application of all the phases are depending on the nature of work.

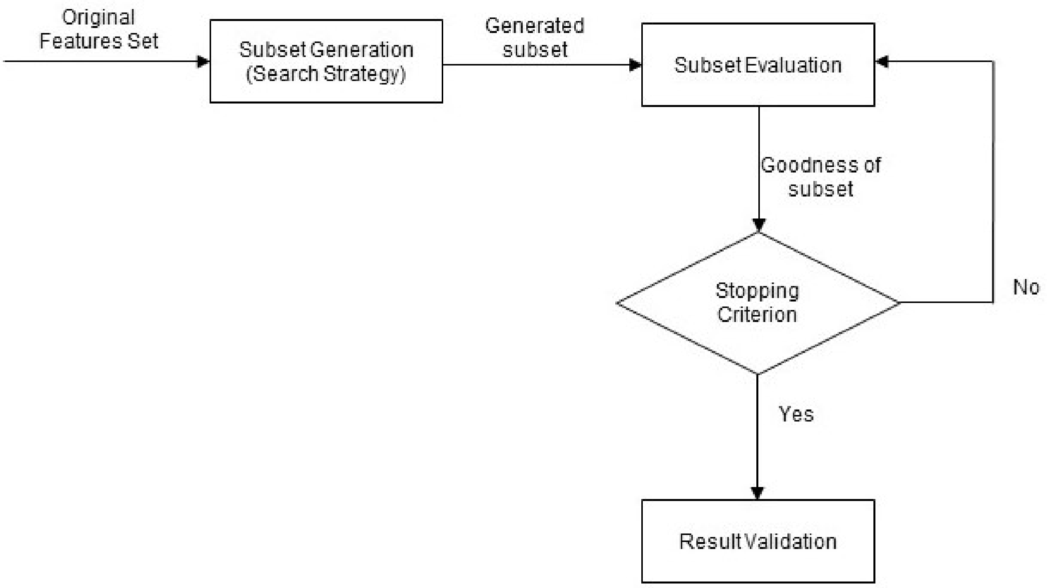

Figure 2.2: General process of feature selection

2.3.1 Subset Generation

In subset generation, which refers to the heuristic process, every search space state is evaluated using a certain subset based on two aspects. Initially, the starting point(s) of the search must be decided, which can affect search direction (Zawbaa et al., 2018; Huan Liu, 2004). It may begin with a full set that discards features or an empty set that gets filled or both discards and fills from both ends concurrently (Huan Liu, 2004;

Molina et al., 2002). The search can also start with a randomly selected subset to prevent from getting trapped in local optima (M. M. Mafarja et al., 2017; Mahajan et al., 2016). The local optima are a problem which typically converges towards a local optimum during the search process for an optimum subset of features. In this step, one must decide on the search strategy (M. M. Mafarja et al., 2017; Mahajan et al., 2016). Usually, for a dataset with N features, 2^N candidate subsets exist. This search space is exponentially extortionate for exhaustive search with even moderate N features (Huan Liu, 2004). The three generic strategies are complete search, heuristic search, and random search.

2.3.1(a) Complete Search

The generation steps perform a complete search in determining the optimal subset based on evaluation criteria. This means; before the final selection, all 2^n possible subset found in search space should be produced and assessed (Huan Liu, 2004). Nevertheless, this complete search may dismiss the idea of being exhaustive. The search space can be reduced by using several heuristic functions without affecting the probability of identifying an optimum outcome. However, a complete search is only viable for small datasets. Even though the order of the search space is $O(2^n)$, fewer subsets are evaluated. The optimality of the feature subset based on the evaluation function is guaranteed as the procedure can backtrack (Dash & Liu, 1997). Backtracking can be performed via branch and bound (Binney & Sukhatme, 2012) and beam search (Rush et al., 2013) techniques. Here, a complete search is carried out, along with termination of search and some branches if the result is poor.

2.3.1(b) Heuristic Search

In every iteration of this generation procedure, all remaining features that are yet to be selected are considered for selection. This simple process has many variations; however, the generation of subsets is incremental, which is either decreasing or increasing (Huan Liu, 2004; Kumar & Minz, 2014). The order of the search space is $O(N^2)$ or less, where N is the number of features. Many heuristic algorithms have been used. This procedure is simple for implementation and rapid for generating outcomes, mainly because the search space is quadratic for the number of features.

2.3.1(c) Random Search

Algorithms used for such an approach generate a new subset at the iterations randomly. With search space denoted as $O(2^n)$, the techniques involved seek fewer subsets than 2^n based on the possible maximum amount of iteration (Mirjalili, 2015; Huan Liu, 2004). Optimum subsets rely on the availability of resources. Every random generation step requires certain parameter value (Mirjalili, 2015; Huan Liu, 2004). This algorithm avoids getting trapped in heuristic search (local minima) but gains the interdependence of features characteristics of the heuristic search. Assigning optimum parameter values is essential to gain exceptional results.

2.3.2 Subset Evaluation

An optimal subset is always relative to a certain evaluation function. Typically, an evaluation function measures the discriminating ability of a feature or a subset to distinguish the different class labels (Huan Liu, 2004). Many studies have investigated the grouping of subset evaluation categories. Upon summarising these works and the latest developments, the evaluation functions can be divided into five categories: distance, information, dependence, consistency, and classifier error rate. Elaboration of these evaluation functions is as follows.

2.3.2(a) Distance Measure

The distance measure is known as separability. It assumes that instances from a varied class are distant in the related features. To address dual-class issue, it is common to prefer feature *X* to *Y*, if *X* induces vast variance between the dual-class conditions than *Y* does; *X* and *Y* turn vague when distance is zero (Wang & Xin, 2005; Huan Liu, 2004). One algorithm used for distance measure is Euclidean distance. Euclidean distance is used by several algorithms to calculate the distance formed between the instances (Xuecheng, 1992; Huan Liu, 2004). Euclidean distance is an algorithm used to measure dependence. The Euclidean distance measures the distance between two points in Euclidean space (Lee et al., 2014). Theoretically, the Euclidean distance algorithm works as follows: for each cell, the distance to each source cell is determined by calculating the hypotenuse with *x_max* and *y_max* (Wang & Xin, 2005; Lee et al., 2014). The shortest distance to a source is determined, and if it is less than the specified maximum distance, the value is assigned to the cell location on the output raster (Lee et al., 2014). The actual algorithm computes the information using a two-scan sequential process. This process makes the speed of the tool independent from the number of source cells, the distribution of the source cells, and the maximum distance specified (Lee et al., 2014).

2.3.2(b) Information Measure

These measures determine the data gain from a feature. The term signifies the extent a feature segregates the instances based on target classification. The statistical measure can be adopted to compare and choose the features. The information gained from a feature *X* is defined as the difference between the prior uncertainty and expected posterior uncertainty using (Huan Liu, 2004; Krishnamurthy & Moore, 1993). The preference is *X* over *Y* if data gained from *X* are more than *Y* (Huan Liu, 2004; Krishnamurthy & Moore, 1993). A commonly used algorithm for information measure is the Shapley Value. It collaboratively seeks the feature contribution to address drawbacks of standard filter techniques (Essen & Wooders, 2018; Shapley, 1953). Shapley Value is defined by Shapley (1953), which offers a method of calculating the contribution given by each player in a coalition setting (Essen & Wooders, 2018). Shapley Value determines power allocation between players in a game of voting (Essen & Wooders, 2018; Rabin, 1993). In Shapley Value, interactions that occur between the players are weighed in to identify the link established between varied players. However, Shapley Value has high computation intricacy (Xin Sun et al., 2013).

2.3.2(c) Dependence Measure

Correlation or dependence determines the capability of estimating a variable value. One commonly used dependence measure refers to the coefficient in seeking the link of a class with a feature (Schlather, 2003). When the link between feature *X* and class *C* exceeds that of features *Y* and *C*, feature *X* is preferred (Huan Liu, 2004). The disparity of this is to determine the dependence of a feature on other features; this value indicates the degree of redundancy of the feature (Schlather, 2003; Huan Liu, 2004).

2.3.2(d) Consistency Measure

The subset inconsistency rate is calculated by weighting upon only the subset features. This rate refers to the number of instance pairs that share similar feature values from varied classes. These measures differ characteristically from the rest due to heavy reliance on the training dataset, as well as the application of Min-Features bias to choose a subset (Almuallim & Dietterich, 1994). Such bias prefers hypotheses with fewer features (Huan Liu, 2004). As a result, a small-sized subset is determined to adhere to the consistency rate.

2.3.2(e) Classifier Error Rate Measure

In this evaluation function, the wrapper approach is used. Since the selection of features is based on a classifier to estimate class labels of unseen instances, the level of accuracy becomes exceptionally high despite expensive computation (Solorio-Fernández et al., 2019; Kohavi & John, 1997).

2.3.3 Stopping Criteria

The halt of the feature selection process is dictated by stopping criteria when: a) search is complete, b) certain bound is attained, where abound can be a specified number, c) further inclusion or exclusion of feature has no impact on the subset, and d) identification of the desired subset is accomplished (Mahajan et al., 2016; Huan Liu, 2004).

2.3.4 Validation

The validation process happens after the feature selection process is completed. It determines the aspect of validity in the selected features via various experiments and later compares with results generated by other techniques (Solorio-Fernández et al., 2019; Huan Liu, 2004; A. Jain & Zongker, 1997). The validation process is usually a straightforward way that measures the outcomes directly with prior knowledge regarding the data (Huan Liu, 2004). With the availability of relevant features beforehand, then they are compared with the chosen features. In most cases where prior knowledge is absent, some indirect techniques are employed to determine changes in mining performance with the features selected.

2.4 The Feature Selection Approaches

Feature selection approaches aids to create an accurate predictive model. These approaches choose features that will give good or better accuracy whilst requiring less data ((I. Jain et al., 2018); (Kumar & Minz, 2014)). Feature selection approaches can be used to identify and remove unneeded, irrelevant, and redundant attributes from data that do not contribute to the accuracy of a predictive model or may decrease the accuracy of the model ((I. Jain et al., 2018); (Kumar & Minz, 2014)). There are three feature selection approaches namely the a) filter, b) wrapper and c) hybrid (Zawbaa et al., 2018; Huan Liu, 2004). The following subsections discuss these approaches.

2.4.1 Filter Approaches

The filter approach is applied in order to access the nominated solution based on the feature's intrinsic attributes (Y. Zhang et al., 2019; H. Liu & Motoda, 2007). As the name says, irrelevant features are filtered out prior to learning based on a certain metric(s). The filter approach does not rely on learning algorithms but examines the relevancy of features by only assessing properties that are integral (X. Sun et al., 2012). Figure 2.3 shows the overall process flow of the filter approach. The filter approach will accept the original features and later a set of features are generated with a measurement metric. The generated set of features will use a stopping criterion before being tested by the learning algorithm before the final set of features is selected. The filter approach is a learning algorithm independent hence making them faster and computationally cheaper than other approaches.

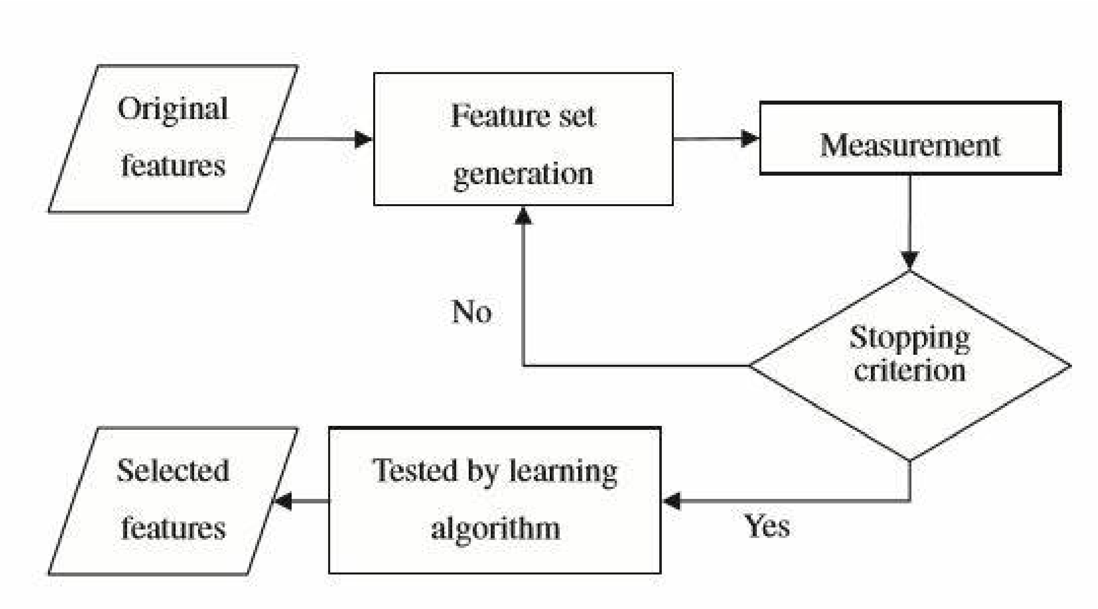

Figure 2.3: The filter approach (Hsu et al., 2011)

In another work by Xin Sun et al. (2012), the feature selection issue was optimised via cooperative game theory. The authors used the Shapley Value proposed by Shapley (1953) in the game theory. This method gave rather interesting outcomes for coalition games in measuring power possessed by the players. The Shapley Value was employed to assess each feature weight. The proposed method used Support Vector Machine (SVM) as a classifier. This method was tested using six real-world datasets derived from the UCI machine learning data repository (Dua & Graff, 2017). The datasets used are KDD Synthetic, Optical recognition, Musk (version 2), Multi-feature pixel, Arrhythmia, and Isolet. The proposed method is evaluated based on the classification accuracy, the number of selected features, and other state-of-art methods selected by the authors. The method was evaluated using three other evaluation criteria: MRMR, Symmetric Uncertainty (SU), and ReliefF. The KDD Synthetic dataset obtained 94.67% with 20 features selected. While the Optical recognition obtained 97.65% with 20 features selected. The Musk (version 2) obtained 95.21% with 26 features selected. Multi-feature pixel on the other hand obtained 89.70% with 25 features selected. The arrhythmia dataset achieved 75.00% with 24 features. The Isolet dataset obtained 76.20% with 30 features. The proposed method gave more exceptional performance than the other evaluation metrics upon choosing several features. The accuracy of the proposed method was the lowest among other evaluation criteria when the first feature was selected, and this continued until the fourth criteria were selected. The proposed method displayed the best accuracy against other evaluation metrics upon incorporating the seventh feature in most of the dataset. Although the proposed approach did not guarantee to retain all useful interdependent/whole interdependent groups, the proposed method gave an effective way to retain useful interdependent features and groups as many as possible. The proposed method did not achieve competitive results on the multi-class classification.

A study by Xin Sun et al. (2012) using cooperative game theory to optimize the feature selection problem. The authors used six real-world datasets from the UCI machine learning data repository (Dua & Graff, 2017). The datasets used are KDD Synthetic, Optical recognition, Musk (version 2), Multi-feature pixel, Arrhythmia, and Isolet. The authors have used the cooperative game theory that employed Shapley Value and mRMR known as CGFS-mRMR and the cooperative game theory that employed Shapley Value and Symmetric Uncertainty (SU) known as CGFS-SU. The methods were evaluated against three other filter-based evaluation criteria; mRMR, SU, and ReliefF. The proposed methods are also evaluated based on the classification accuracy and number of selected features. The CGFS-mRMR obtained the following results. The KDD Synthetic dataset obtained 96.50% with 12 features selected. While the Optical recognition obtained 98.72% with 19 features selected. The Musk (version 2) obtained 95.56% with 28 features selected. Multi-feature pixel on the other hand obtained 90.50% with 20 features selected. The arrhythmia dataset achieved 76.10% with 23 features. The Isolet dataset obtained 81.98% with 30 features. The CFGS-SU obtained the following KDD Synthetic dataset obtained 95.67% with 13 features selected. While the Optical recognition obtained 98.24% with 20 features selected. The Musk (version 2) obtained 95.92% with 28 features selected. Multi-feature pixel on the other hand obtained 84.75% with 24 features selected. The arrhythmia dataset achieved 75.33% with 18 features. The Isolet dataset obtained 74.98% with 28 features. The interdependent groups of features commonly exist in the traditional feature selection algorithms that always do not retain the useful intrinsic groups. The experimental results from the proposed methods show the improved performance of representative feature selection.

Zhang et al. (2011) introduced mRMR optimized classification for automatic glaucoma diagnosis. The dataset used in this work is from ORIGA (Almazroa et al., 2018).