FUZZY BASED META-HEURISTIC APPROACHES FOR ATTRIBUTE REDUCTION PROBLEM

MAJDI M. N. MAFARJA

THESIS SUBMITTED IN FULFILMENT FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2012

PENDEKATAN META-HEURISTIK BERASASKAN KABUR UNTUK
MASALAH PENGURANGAN ATRIBUT

MAJDI M. N. MAFARJA

TESIS YANG DIKEMUKAKAN UNTUK MEMPEROLEH IJAZAH
DOKTOR FALSAFAH

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2012

DECLARATION

I hereby declare that the work in this thesis is my own except for quotations and summaries which have been duly acknowledged.

3rd September 2012

MAJDI M. N. MAFARJA
P51681

ACKNOWLEDGMENTS

All the praises be to the mighty Allah, the Merciful and the Beneficent for the strength and blessing in the completion of this study.

Indeed there are many wonderful people who have contributed significantly throughout the whole course of my study up to the completion of this thesis. I owe a great deal to them.

First and foremost, I wish to express my most sincere acknowledgment to my supervisor: Assoc. Prof. Dr. Salwani Abdullah for her valuable guidance, generosity and freedom throughout the entire research and thesis writing.

To my dearest mother and father, thank you for bringing me up to who I am today. My success symbolizes and reflects the support and love from both of you. My deepest appreciation goes to my wife for her love, patience and understanding.

My thanks to my friend Dr. Hamza Turabieh, who helped me to be a student in UKM under the supervision of Dr. Salwani. My thanks also to my friend Nasser R. Sabar for his help and support.

I am very grateful to UKM and all the staff and members of the School of Computer Science for the help, and friends who have assisted me whenever I need them.

ABSTRACT

Attribute reduction represents a NP-hard problem that can be defined as the problem of locating a minimal subset of attributes from an original set. The key issue associated with feature selectors is the production of a minimal number of reducts that represents the original meaning of all features. Rough Set Theory has been used for attribute reduction with much success. This is due to the fact that rough set theory uses only the supplied data during the *feature selection* process and no more information is needed. The reduction method inside rough set theory is applicable only to small data sets because finding all possible reducts is a time-consuming process. However, no approach can ensure optimality when solving this problem. Some approaches are more efficient than others due to some of the characteristics of the algorithm such as the number of parameters involved. The aim of the research presented in this thesis is to provide effective approaches for finding the most informative and minimal attributes with least information loss. This has been achieved via a number of meta-heuristic approaches which mainly depend on two algorithms, i.e., the Record-to-Record Travel algorithm and the Great Deluge algorithm. Both algorithms are deterministic optimisation algorithms whose structures are inspired by and resemble the Simulated Annealing algorithm but differ in the acceptance of worse solutions. Moreover, they belong to the same family of meta-heuristic algorithms that are used to avoid the local optima by accepting non-improving neighbours. The research first highlights the use of the record-to-record travel algorithm in solving attribute reduction problem, and then examines the effects of enhancing the algorithm by incorporating a Fuzzy Logic Controller in order to intelligently control the parameter involved in the algorithm (called the Fuzzy Record-to-Record Travel algorithm). Next, two modifications of the Great Deluge algorithm are investigated, where the search space is divided into three regions. Instead of using a linear mechanism to update the water level (as in the original Great Deluge algorithm), the modified Great Deluge algorithm updates the water level for each region using a different scheme which is based on the quality of the trial solution. Then, a fuzzy logic controller is used to control the updated scheme of the water level (called the Fuzzy Great Deluge algorithm). This research further investigates the efficacy of the hybridisation approach between the aforementioned algorithms with the Genetic Algorithm (called Fuzzy Record-to-Record Travel with Genetic Algorithm and Fuzzy Great Deluge with Genetic Algorithm). Experimental results show that the fuzzy Great Deluge with Genetic Algorithm approach outperforms the other proposed approaches here and is effective for most of the University of California Irvine benchmark data sets when compared to other available approaches in the literature.

ABSTRAK

Pengurangan atribut merupakan masalah NP-sukar yang ditakrifkan sebagai masalah pencarian subset atribut minimal daripada set asal. Isu utama yang dikaitkan dengan pemilih fitur adalah untuk menghasilkan pengurangan minimum yang mewakili makna asal semua fitur. Teori set kasar telah diguna dengan jayanya dalam pengurangan atribut kerana teori set kasar menggunakan hanya data yang dibekalkan semasa proses pemilihan fitur dan tiada maklumat lanjut diperlukan. Kaedah pengurangan dalam teori set kasar hanya sesuai untuk set data yang kecil kerana proses mendapatkan semua kemungkinan pengurangan atribut memerlukan masa yang panjang. Walau bagaimanapun, tiada pendekatan yang dapat memastikan penyelesaian optimal dalam masalah ini. Beberapa pendekatan lebih cekap daripada pendekatan yang lain kerana ciri-ciri yang ada pada pendekatan tersebut seperti bilangan parameter yang diperlukan olehnya. Tujuan penyelidikan ini adalah untuk menyediakan pendekatan yang efektif untuk mencari pengurangan atribut yang minimum dan bermaklumat dengan kehilangan maklumat yang sedikit. Ini dicapai melalui dua pendekatan meta-heuristik iaitu algoritma Rekod-Rekod Perjalanan dan algoritma Banjir Besar. Kedua-dua algoritma tersebut adalah algoritma pengoptimuman berketentuan yang diinspirasi dan menyerupai struktur algoritma Simulasi Penyepuhlindapan manakala berbeza dalam penerimaan penyelesaian yang kurang baik. Selain itu, mereka tergolong dalam keluarga algoritma meta-heuristik yang sama yang digunakan untuk mengelakkan optima tempatan dengan menerima kejiranan yang tidak-meningkat. Penyelidikan yang pertama menekankan penggunaan algoritma Rekod-Rekod Perjalanan dalam menyelesaikan masalah pengurangan atribut, dan mengkaji kesan penggabungan pengawal logik kabur untuk mengawal dengan pintar parameter yang terlibat dalam algoritma tersebut (dikenali Rekod-Rekod Perjalanan Kabur). Seterusnya, dua pengubahsuaian terhadap algoritma Banjir Besar diselidiki dengan membahagikan ruang carian kepada tiga kawasan. Di setiap kawasan, paras air dikemas kini menggunakan skema yang berbeza berdasarkan pada kualiti penyelesaian percubaan dan bukannya menggunakan mekanisma linear seperti dalam algoritma Banjir Besar asal (dikenali modifikasi Banjir Besar). Pengawal logik kabur kemudiannya digunakan untuk mengawal skema paras air yang dikemaskini (dikenali Banjir Besar Kabur). Akhir sekali, penyelidikan terhadap keupayaan penghibridan antara algoritma yang disebutkan di atas dengan Algoritma Genetik (dikenali Rekod-Rekod Perjalanan Kabur dengan Algoritma Genetik dan Banjir Besar Kabur dengan Algoritma Genetik). Prestasi setiap pendekatan diuji pada set data piawai UCI. Keputusan eksperimen menunjukkan pendekatan Rekod-Rekod Perjalanan Kabur dengan Algoritma Genetik lebih baik daripada pendekatan lain yang dicadangkan dalam kajian ini dan efektf pada kebanyakan set data yang diuji apabila dibandingkan dengan pendekatan lain yang terdapat dalam kesusasteraan.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Acronym	**Definition**
ACO	Ant Colony Optimization
AIS	Artificial Immune Systems
AR	Attribute Reduction
BC	Bee Colony
CHH	Constructive Hyper-heuristics
CNS	Composite Neighbourhood Structure
FLC	Fuzzy Logic Controller
FuzzyGD	Fuzzy Great Deluge Algorithm for Attribute Reduction
FuzzyRRT	Fuzzy Record to Record Travel Algorithm for Attribute Reduction
GA	Genetic Algorithm
GA-FuzzyGD	Fuzzy Great Deluge Algorithm with Genetic Algorithm
GA-FuzzyRRT	Fuzzy Record-to-Record Travel Algorithm with Genetic Algorithm
GDA	Great Deluge algorithm
HVNSA	Hybrid Variable Neighbourhood Search algorithm
IDS	Intelligent Dynamic System
ILS	Iterated Local Search
KDD	Knowledge Discovery in Databases
m-GD	Modified Great Deluge Algorithm for Attribute Reduction
NP	None polynomial
PSO	Particle Swarm Optimisation

RRT	Record-to-Record Travel
RST	Rough Set Theory (RST)
SA	Simulated Annealing
Sol_{best}	The Best Solution
Sol_{trial}	The Trial Solution
SS	Scatter Search
TA	Threshold Accepting
TS	Tabu Search
UCI	University of California Irvine
VNS	Variable Neighbourhood Search

CHAPTER I

INTRODUCTION

1.1 BACKGROUND AND MOTIVATION

The on-going explosive growth of real data has led to a great challenge in the field of services and solution tools development. Central to this issue is the knowledge discovery process, particularly Knowledge Discovery in Databases (KDD) (Fayyad et al. 1996b). Knowledge discovery in databases is concerned with extracting useful knowledge from databases (Fayyad et al. 1996a). The means of meeting the challenge of extracting knowledge from databases draws upon research in many areas such as statistics, data visualisation, pattern recognition, optimisation, machine learning, and high-performance computing. Fayyad et al. (1996b) define KDD as *"the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data"*.

Based on the study by Fayyad et al. (1996a), the KDD process can be decomposed into five steps illustrated in Figure 1.1: (i) data selection where a data set is selected; (ii) data cleaning/preprocessing, which includes noise removal or reduction, missing value imputation, and attribute discretisation; (iii) data reduction, which aims to find the most informative features from a data set by removing the redundant and irrelevant features that will not aid KDD and may in fact mislead the process; (iv) data mining, which deals with extracting the hidden predictive information from large databases according to the goals of the knowledge discovery task; and (v) evaluation, which mainly check the validity, usefulness, novelty and simplicity of the discovered knowledge. This process may require the repetition of some of the previous steps.

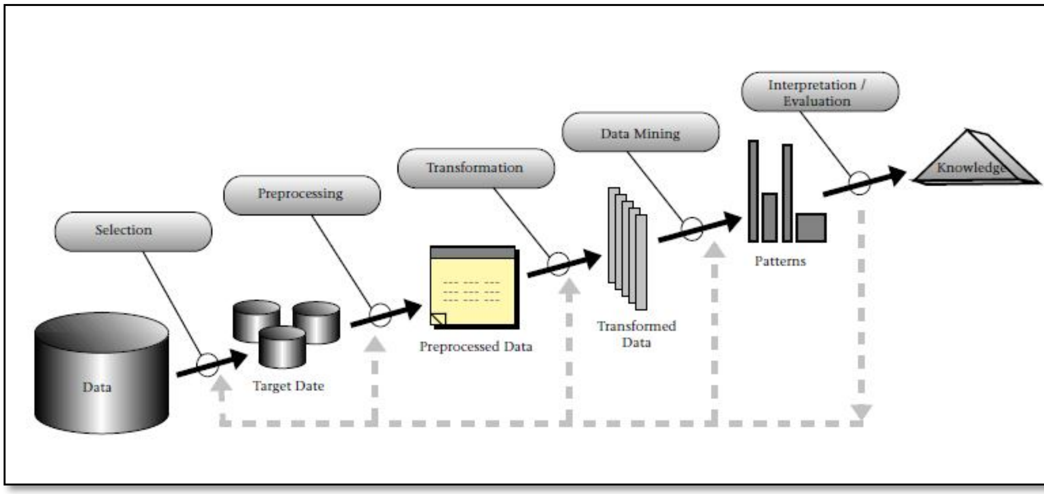

Figure 1.1 Knowledge discovery process steps

Source: (Fayyad et al. 1996a)

The third step in the knowledge discovery process, namely data reduction, is the point of interest for this research. Attribute Reduction (AR) (or data reduction) is regarded as an important preprocessing technique in machine learning and in the data mining process (Hu et al. 2010; Zhang et al. 2008). It can be defined as the problem of finding a minimum reduct (subset) from the original set. Liu and Motoda (1998) define attribute reduction as *"a process that chooses an optimal subset of features according to certain criterion"*.

The attribute reduction process aims to eliminate the irrelevant and redundant attributes from high-dimensional data sets which increase the chances that a data mining algorithm will find spurious patterns that are not valid in general. Furthermore, when dealing with high-dimensional data sets, a longer time is needed to find the desired results. Liu and Motoda (1998) indicate that the purposes of AR are to: (a) improve the performance (speed of learning, predictive accuracy, or simplicity of rules); (b) visualise the data for model selection; and (c) reduce the dimensionality and remove the noise.

Rough Set Theory (RST) (Pawlak 1982) is considered an effective mathematical tool for dealing with uncertain, imprecise and incomplete information. It has been successfully applied in such fields as knowledge discovery, decision support, and pattern classification (Jensen 2005). Attribute reduction is a key problem in RST, and finding a minimal attribute reduction has also been described as a NP-hard problem (Komorowski et al. 1999). Minimal attributes can be determined by two main approaches, i.e., discernibility functions-based and attribute dependency-based approaches (Han et al. 2003; Hoa 1996). However, these approaches have a drawback because of their intensive computations either of discernibility functions in the former or of positive regions in latter (Han et al. 2005).

Many researchers are focusing on the problem of finding a subset with minimal attributes from an original set of data in an information system (Anaraki & Eftekhari 2011; Deng et al. 2012; Kabir et al. 2012; Kabir et al. 2011; Kaneiwa & Kudo 2011; Liang et al. 2012; Liu et al. 2009; Liu et al. 2002; Suguna & Thanushkodi 2010; Swiniarski & Skowron 2003). Locating such a subset is basically done by using the complete search approach, which locates all possible subsets to find the optimal subsets (i.e., those with a maximum rough set dependency degree). Obviously, a complete search approach is an impractical and complex solution to the problem and is only practical for simple data sets. For high-dimensional data sets, the heuristic search is much faster than the complete search, because it searches according to a particular path in order to find the minimal reduct (Liu & Motoda 1998).

An alternative way to determine a minimal reduct is to adapt meta-heuristic algorithms. Meta-heuristics can be classified into two families of methods, i.e., single-solution-based methods and population-based methods. On the one hand, examples of single-solution-based methods are: Tabu Search, Simulated Annealing, Threshold Accepting, Variable Neighbourhood Search, Iterated Local Search, Guided Local Search and the Greedy Adaptive Search Procedure. On the other hand, population-based meta-heuristics methods include evolutionary algorithms (Genetic Algorithms, Evolution Strategies, Genetic Programming, Evolutionary Programming, Estimation of Distribution Algorithms, Differential Evolution, and Co-evolutionary Algorithms), swarm intelligence-based methods (e.g., Ant Colony, Particle Swarm Optimisation), Scatter Search, Bee Colony and Artificial Immune Systems (Glover & Kochenberger 2003; Talbi 2009).

In the literature, many meta-heuristic-based methods which were designed to solve the attribute reduction problem can be found, such as the Genetic Algorithm (Handels et al. 1999; Jensen & Shen 2004; Wroblewski 1995), Particle Swarm Optimisation (Wang et al. 2007b), Ant Colony algorithm (Jensen & Shen 2003; Ke et al. 2008), Tabu Search (Hedar et al. 2006), Great Deluge algorithm (Abdullah & Jaddi 2010), Composite Neighbourhood Structure (Jihad & Abdullah 2010), Hybrid Variable Neighbourhood Search algorithm (Arajy & Abdullah 2010) and Constructive Hyper-Heuristics (Abdullah et al. 2010a), Bees Algorithm (Alomari & Othman 2012).

In these methods, the parameters are manually tuned through a series of preliminary experiments in advance before they are employed to handle the given problems. Manual parameter tuning has some drawbacks because the parameters are sometimes independent, thus trying all possible combinations of the parameters is a time-consuming process. Furthermore, it is not necessary that the parameters' values are optimal because the given data sets are applicable to other data sets even if the effort made in setting them was significant (Eiben et al. 1999). This motivates the work in this thesis to propose a good parameter-tuning method that helps in finding good parameter settings for a given problem.

Since the performance of the heuristic approaches can vary significantly with different parameters values (Dueck 1993; Gendreau 2003) and may be improved after changing the parameters' values (Glover & Kochenberger 2003), this opens a research opportunity to further investigate the effects of controlling the parameters during the search process. Therefore, unlike previous studies, this research investigates possible meta-heuristic approaches that may be improved by employing a Fuzzy Logic Controller (FLC) to intelligently control the parameters involved (in each proposed method) in tackling the attribute reduction problem.

Moreover, Hart et al. (2005) observe that each meta-heuristic algorithm can perform in a different way according to the problem in hand, thus, no significant solver can be applied successfully and efficiently for all purposes. Since each algorithm has specific strengths and weaknesses, this research is motivated to investigate a hybridisation approach that combines a population-based approach with a single-solution-based approach in order to gain the benefits of the exploration and exploitation mechanisms offered by population-based and single-based approaches, respectively. Through achieving a balance between exploration and exploitation throughout the searching process, it is believed that it is possible to generate better solutions for the problem in hand.

1.2 PROBLEM STATEMENT

There is on-going research interest in attempting to find better solutions for attribute reduction problems (Abdullah et al. 2010a; Dorigo & Blum 2005; Jensen & Shen 2003; Jensen & Shen 2004; Ke et al. 2008; Wang et al. 2007a; Xu et al. 2009). Due to the complexity of the problem, the modelling of such a problem is quite complex, and mathematical searching (e.g., by using exact methods) for an optimal solution is usually impractical in terms of computational times. The use of RST has proved successful in achieving data reduction (Bello et al. 2009; Jensen 2005). This success is due in part to the fact that no additional information about the data is needed, and it only analyses the hidden facts in the data (Jensen 2005). Many meta-heuristic approaches have been applied to solve the problem of attribute reduction in RST, such as Simulated Annealing (Jensen & Shen 2004), Tabu Search (Hedar et al. 2006), Genetic Algorithm (Jensen & Shen 2004), Particle Swarm Optimisation (Wang et al. 2007b), and Ant Colony (Jensen & Shen 2003; Kabir et al. 2012; Ke et al. 2008; Ming 2008), Dynamic Mesh Optimization (Bello et al. 2008), Fire Fly algorithm (Banati & Bajaj 2011), Electromagnetism-like Mechanism (Su & Lin 2011),Wasp Swarm Optimization (Fan & Zhong 2012). However, no approach can ensure optimality, and some approaches are more efficient than others.

Although there are a vast number of methods available to find minimal attributes, there is still a great interest in proposing new algorithms. The motivation for this is twofold: first, determining the minimal attributes produced by the existing attribute reduction algorithms can still be improved. Second, the classification accuracy obtained by using the minimal attributes that are produced by the existing algorithms can still be bettered. Although many meta-heuristic approaches have been applied successfully on attribute reduction, the performance of each meta-heuristic crucially depends on its parameter(s) (Gendreau 2003). Therefore, in this work, a fuzzy logic controller is chosen to intelligently control the parameter values of the proposed meta-heuristic methods during the search process with the aim of further improving the performance of these methods.

This research proposes a number of meta-heuristic approaches which mainly depend on two algorithms, i.e., the Record-to-Record Travel (RRT) algorithm and the Great Deluge (GD) algorithm. The significance of these methods relates to the ease of their implementation and the number of required parameters, which influences the performance of the algorithms (Dueck 1993). Furthermore, both of these algorithms are deterministic optimisation algorithms, and their structures are inspired by and resemble the Simulated Annealing (SA) algorithm, while they differ in the acceptance of worse solutions. When Dueck (1993) compared the performance of SA, GD and RRT, he found that both GD and RRT performed better than SA. Moreover, they belong to the same family of meta-heuristic algorithms that are used to avoid the local optima by accepting non-improving neighbours or solutions. Furthermore, this research also studies the effect of embedding a fuzzy logic controller to intelligently control the parameters involved in the individual approaches and later hybridises these approaches with the addition of a population-based algorithm (in this case, the Genetic Algorithm).

In the context of finding the minimal attributes from an original data set using meta-heuristic approaches, this thesis seeks to answer the following three research questions:

i. Does the intelligent mechanism of controlling the parameters involved in single-based meta-heuristic methods increase the probability of finding high-quality minimal attributes?

ii. Does the hybridisation between a population-based and a single-based approach obtain better results compared to a single-based approach in isolation?

iii. What is the relationship between attribute reduction techniques and the resulting classification accuracy and the number of generated rules?

1.3 RESEARCH OBJECTIVES

The main aim of this research is to propose efficient approaches that are able to produce good-quality solutions for the attribute reduction problem. Based on this research, it is expected to achieve the following objectives which will contribute to the literature on the attribute reduction problem. The pertinent objectives are threefold:

i. To propose a fuzzy-based meta-heuristic approach that intelligently controls the parameter of the algorithm that involved in solving the attribute reduction problem;

ii. To propose a hybrid approach (between Genetic Algorithm and fuzzy single-based approaches) to increase the ability of exploration and exploitation in finding minimal attributes; and

iii. To investigate the classification accuracy and the number of generated rules by using the selected attributes from the proposed algorithms.

1.4 RESEARCH SCOPE

As mentioned above, attribute reduction is a very important process that is used to enhance the performance of data mining tasks. This research is concerned with developing alternative meta-heuristic methods (in this work, fuzzy-based meta-heuristic methods) to tackle this problem and to enhance the performance of these approaches through hybridisation with a Genetic Algorithm (GA) in order to gain the respective benefits of each approach.

In this work, 13 well-known University of California Irvine (UCI) data sets, which have been used by many researchers in the literature (Abdullah & Jaddi 2010; Hedar et al. 2006; Jensen & Shen 2004; Wang et al. 2007a), are used to test the performance of the proposed algorithms. These data sets can be freely downloaded from http://www.ics.uci.edu/~mlearn/ (Blake & Merz 1998). The performance of the proposed algorithms is evaluated based on the minimal number of selected attributes, the classification accuracy and the number of rules generated by using the selected attributes.

1.5 ATTRIBUTE REDUCTION METHODS

Two main taxonomies for attribute reduction methodologies have been proposed. The first taxonomy was introduced by Dash and Liu (1997), who considered attribute reduction as a search problem. They classified each attribute reduction method by the employed searching method. The second taxonomy was suggested by Langley (1994), who grouped different attribute reduction methods into two broad groups: filters and wrappers.

1.5.1 Data Preprocessing

Data preprocessing is the second phase in the KDD process, the aim of which is to get some idea about the data and prepare it for the next step in the process (Liu & Motoda 1998). It is an important phase in the KDD process not least because of the increasing amount of data that this process has to contend with. According to Liu and Motoda (1998), the data preprocessing phase includes: target data selection, data cleaning, and data projection and reduction. Target data selection creates a data set. The data cleaning process deals with filling-in missing values, reducing noisy data, detecting and eradicating data deviation and solving discrepancies. Data projection and reduction deals with reducing the size of data sets and transforming the data into a proper form as per the needs of the application.

Advances in data collection and storage capabilities during the past few decades have led to attribute reduction becoming a useful and indeed sometimes necessary tool in many fields (such as in pattern recognition and machine learning) that can be used to reduce the data dimensionality to a more manageable size with as little information loss as possible (Fodor 2002; Jensen 2005).

1.5.2 Attribute Reduction as a Search Problem

According to Dash and Liu (1997), in a typical attribute reduction method there are four basic steps (see Figure 1.2), i.e., (a) a generation procedure to generate the next candidate subset: (b) an evaluation function to evaluate the generated subset; (c) a stopping criterion to decide when to terminate the process; and (d) a validation procedure to check the validity of the subset.

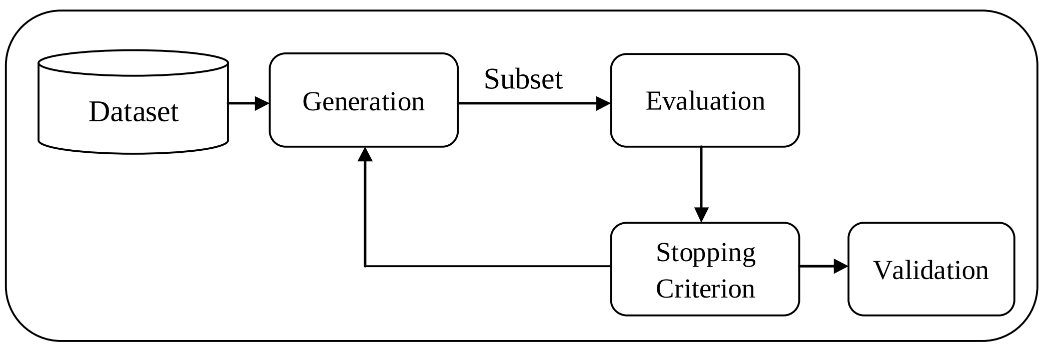

Figure 1.2 Attribute reduction process with validation

Source: (Jensen 2005)

a. Subset generation

Subset generation is a search procedure (Langley 1994; Siedlecki & Sklansky 1988). Basically it selects a subset of features for evaluation. The process may start with no features where features are added (forward selection), or start with all features and then features are removed (backward elimination), or with a random feature subset where features are either iteratively added or removed or produced randomly thereafter (Langley 1994). Each subset of the possible feature represents a state in the search space. In a data set with N features, there should be a total of 2^N subsets. In the case of three features there are eight subsets (states), as shown in Figure 1.3 (Liu & Motoda 1998).

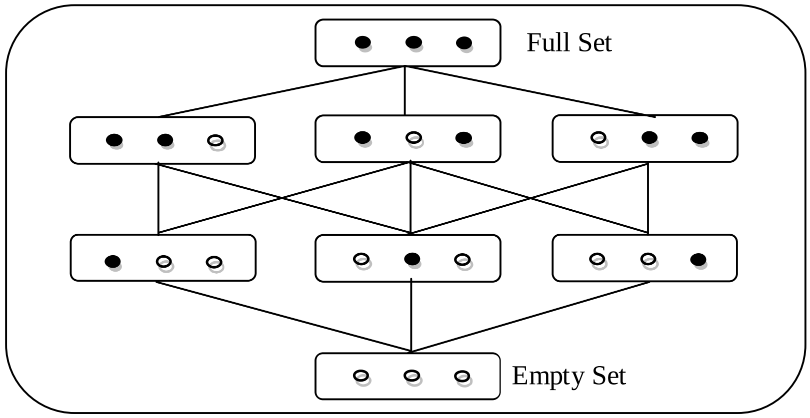

Figure 1.3 Attribute reduction as a search problem

In Figure 1.3, the first state (full set) represents the full subset where three features are selected, while the other state (empty set) represents the empty subset in which no feature is selected. The generation procedure selects a set of features from all the attributes of a sample using one of three search strategies: a complete search, a heuristic search, or a random search (Dash & Liu 1997).

The complete search will produce all possible subsets and find the best ones. It is the most computationally intensive method because the space complexity (the number of subsets to be generated) is $O(2^N)$. The complete search method either starts with an empty set or a full set and it progresses through the features (one feature at a time) until all features have been evaluated and a minimum subset is found. Even though this method is computationally expensive, it ensures that no optimal subset can possibly be missed.

The heuristic search employs a heuristic in conducting a search. The heuristic search can be described as a 'depth first' search guided by heuristics. The cost of this process may be just a path connecting the two ends, as in Figure 1.3, which may take a maximum length of N. The space complexity (the number of subsets to be generated) is $O(N)$. Due to the fact that the heuristic search searches only a particular path, it is obviously faster than the complete search. However, it risks losing optimal solutions.

The random search differs from the previous two search strategies in that, as the name suggests, it selects features randomly. Randomisation can be beneficial to the attribute reduction process in that there is no need to wait until the search ends. However, the random process may result in a solution that is complete, but not necessarily optimal.

Table 1.1 shows the possible combinations of search strategies and search directions. The symbol '√' means that a combination is sensible, while '×' means the opposite.

Table1.1 Search strategies and search directions

Search Direction	Search Strategy		
	Complete	Heuristic	Random
Forward selection	√	√	×
Backward elimination	√	√	×
Bidirectional	√	√	×
Random	×	√	√

Source: (Liu & Motoda 1998)

b. Evaluation function

In all search strategies, the need for an evaluation function is a common issue. It defines the 'goodness' of a feature or a subset of features (Liu & Motoda 1998). Dash and Liu (1997) divided the evaluation functions into five categories, i.e., (i) distance measures, (ii) information measures, (iii) dependence measures, (iv) consistency measures, and (v) classifier error rate measures. The use of different evaluation functions may result in a different optimal subset (Dash & Liu 1997). In this research, the dependence measure is discussed in further detail.

Liu and Motoda (1998) summarise all the possible features selection algorithms in terms of search strategies, search directions and evaluation functions in a three-dimensional structure, as shown in Figure 1.4.

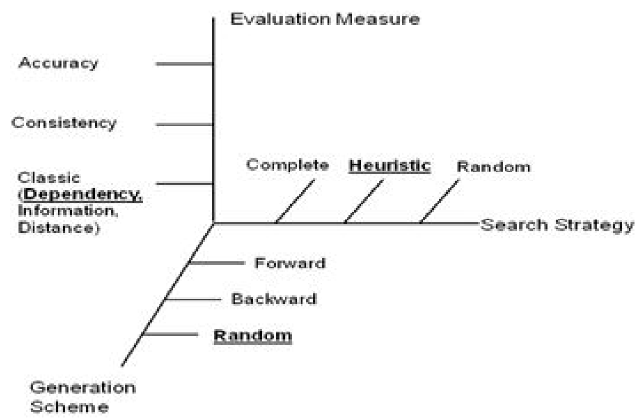

Figure 1.4 Three principal dimensions of attribute reduction: search strategy, evaluation measure, and generation scheme
Source: (Liu & Motoda 1998)

c. Stopping criterion

The stopping criterion terminates the search process when a certain condition is satisfied. The stopping criterion is usually determined by a particular combination of the used searching strategy and the evaluation function (Liu & Motoda 1998). A stopping criterion is based on the searching process and may be a predefined number of selected features or a predefined number of iterations (Dash & Liu 1997).

d. Validation

The validation process usually takes place after the attribute reduction process has finished. This process tests the validity of the selected attributes by carrying out different tests and comparing the results with those that have been produced by other attribute reduction methods.

1.5.3 Filter and Wrapper Attribute Reduction Methods

Langley (1994) and Blum and Langley (1997) considered a different taxonomy than the one presented in Section 2.3.1. They divided the attribute reduction methods into three types of model: filter, wrapper, and embedded. Here, the filter and wrapper models are discussed.

In a filter model, the selection process is performed independently from the induction algorithm. The goodness of a feature or a subset of features is evaluated depending on certain properties of the data (as shown in Figure 1.5). The filter model is computationally cheap when compared with the wrapper model because the former does not involve any induction algorithm (Liu & Motoda 1998). However, filter models may suffer through low performance of the induction algorithm if the selected features do not match induction algorithm. Examples of filter models include Chi-Square (Liu & Setiono 1995), Information Gain (Quinlan 1986), Gain Ratio (Quinlan 1993), ReliefF (Robnik-Šikonja & Kononenko 2003).

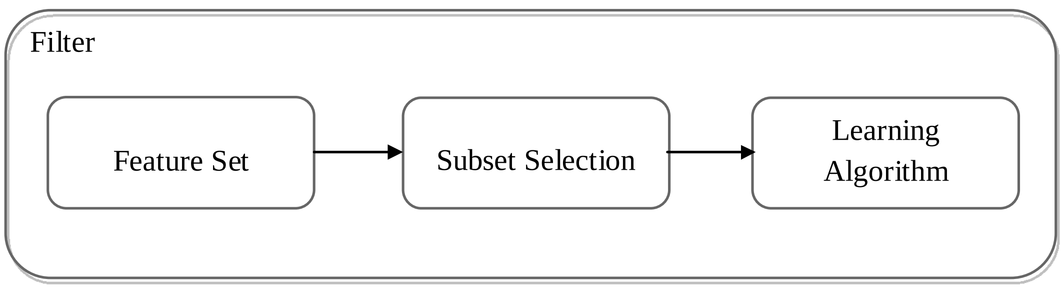

Figure 1.5 Filter attribute reduction

A wrapper model, which is essentially the opposite of a filter model, uses the induction algorithm to directly evaluate the feature subsets (as shown in Figure 1.6). The aim of machine learning algorithms is to achieve the highest performance of a classifier in learning from data (Liu & Motoda 1998). Therefore, the selection of features should consider the characteristics of the classifier as evaluation criteria to achieve higher accuracy. However, because a wrapper model has to train the classifier for each subset evaluation, the complexity and thus the time needed will be much greater than in the case of a filter model (Kaneiwa & Kudo 2011). Due to the fact that each classifier has unique characteristics, it is not wise for a machine to learn a classifier by using attributes selected by another classifier. Examples of wrapper models include the LVW algorithm (Liu & Setiono 1996) and a neural network-based method (Setiono & Liu 1997).

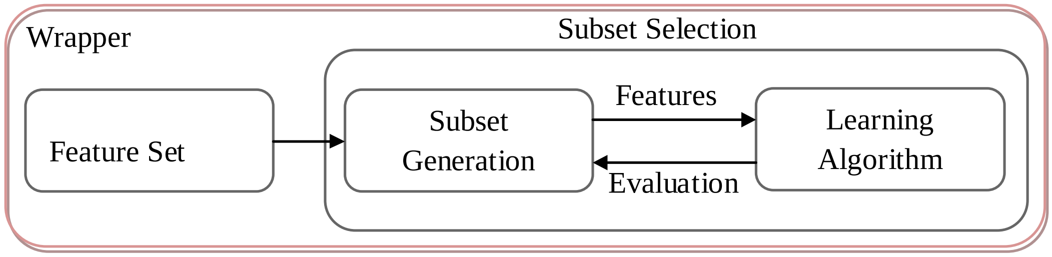

Figure 1.6 Wrapper attribute reduction

Source: (Jensen 2005)

1.6 ROUGH SET THEORY FOR THE ATTRIBUTE REDUCTION PROBLEM

Rough set theory is an effective computing tool that can be used to discover data dependencies and to reduce the number of attributes in a data set, and it requires no additional information. Rough set theory (Pawlak 1991; Pawlak 1982) is a mathematical approach to solve vagueness, imprecision and uncertainty problems. Rough set theory is embedded in the classical set theory which supports approximation in decision making. Rough set concepts can be defined by topological operations such as interior and closure, which are considered to be approximations. These two operations are called the lower and upper approximation. The lower approximation in the domain of knowledge is the set of objects that are known with certainty to belong to the subset of interest. The upper approximation is the set of objects that can possibly be classified as part of the subset of interest. The boundary region is the difference between the upper and the lower approximations.

The starting point of rough set theory is the concept of indiscernibility (Pawlak 1982). Let an information system be $S = (U, A)$, where U is a non-empty set of finite objects called the universe of discourse and A is a non-empty set of attributes. Every attribute $a \in A$, is associated with a set of its values (V_a). Any subset B of A determines a binary relation $IND(B)$ on B, which is called an indiscernibility relation. If $(x, y) \in IND(B)$, then x and y are *B-indiscernible*. The relation $IND(B)$ can be defined as follows:

$$IND(B) = \{ (x, y) \in U^2 | \forall a \in B, a(x) = a(y) \} \quad (1.1)$$

Next, the indiscernibility relation is used to define the upper and lower approximations and the basic concepts of RST. For a subset X $\subseteq$ U, the B-*lower* and B-*upper* approximations can be defined as follows:

$$\underline{B}X = \{x \mid [x]_P \square X\} \quad (1.2)$$

$$\overline{B}X = \{x \mid [x]_P \cap X \neq \varphi\} \quad (1.3)$$

Let D and C be subsets of A, then the positive, negative and boundary regions can be defined as follows:

$$POS_C(D) = \bigcup_{X \square U/I(D)} C \ (X) \quad (1.4)$$

$$NEG_C(D) = U \quad \bigcup_{X \square U/I(D)} C^*(X) \quad (1.5)$$

$$BND_C(D) = \bigcup_{X \square U/I(D)} C^*(X) \quad \bigcup_{X \square U/I(D)} C \ (X) \quad (1.6)$$

The use of rough set theory for attribute reduction has many advantages. The main advantage is that a rough set does not require any preliminary or additional information about the data, such as probability in statistics or basic probability assignment in Dempster–Shafer theory and grade of membership or the value of possibility in Fuzzy Set Theory. To illustrate the concept of a rough set and attribute reduction, an example data set is used (see Table 1.2). Data is often presented as a table, the columns of which are labelled by attributes, while the rows are objects of interest, and entries of the row are attribute values. Here, the example dataset (in Table 1.2) consists of three conditional attributes *C* = {a, b, c, one decision attribute *D* = {d} and six objects. The task of attribute reduction is to determine the minimal reduct from the conditional attributes so that the resulting reduced data set remains consistent with respect to the decision attribute.

Table1.2 Example data set

U	a	b	c	D
1	0	0	0	1
2	0	0	1	0
3	0	0	2	0
4	1	0	0	1
5	1	1	1	1
6	1	0	2	0

As in Table 1.2, it can be stated that when considering a subset B = {a, b}, the objects 1, 2, and 3 certainly belong to a class in attribute {d}, which is indiscernible along with objects 4 and 6. Then, we have the following:

$$U/IND(B) = \{\ \{1,2,3\}, \{4,6\}, \{5\}\ \}$$

An often applied measure in data analysis is the *dependency degree* between attributes (Düntsch & Gediga 1998). Intuitively, a set of attributes D *depends totally* on a set of attributes C, denoted as $C \Rightarrow D$, if all of the values of the attributes from D are uniquely determined by the values of the attributes from C. If there exists a functional dependency between the values of D and C, then D depends totally on C. Dependency can be defined as follows:

For D, $C \subset A$, it is said that D depends on a degree of k ($0 \leq k \leq 1$), denoted as $C \Rightarrow_k D$, if

$$k = \gamma_C(D) = \frac{|POS_C(D)|}{|U|} \tag{1.7}$$

where $|F|$ denotes the cardinality of set F.

If $k = 1$, then D depends totally on C. If $k < 1$, then we can claim that D depends partially on C, and if $k = 0$, then we can say that D does not depend on C. In the example data set in Table1.2, let $C = \{a, b\}$ and $D = \{d\}$. Then, we have the following:

$$U/IND(D) = \{\ \{1,4,5\}, \{2,3,6\}\ \},$$

$$POS_C(D) = POS_{\{a,b\}}(\{d\}) = \bigcup\{\{5\}, \varphi\} = \{5\},$$

$$\gamma_{\{a,b\}}(\{d\}) = \frac{\left|POS_{\{a,b\}}(\{d\})\right|}{|U|} = \frac{1}{6}.$$

One of the major applications of rough set theory is that it is used to determine the minimal attributes by eliminating the redundant attributes from the original sets, without any information loss (Pawlak 1982; Pawlak 1991). The reduction of attributes can be achieved by comparing the dependency degrees of the generated subsets so that the reduced set has the same dependency degree as the original set (Jensen & Shen 2004). A reduct is formally defined as a subset R of the minimal cardinality of the conditional attribute set C such that $\gamma_R(D) = \gamma_C(D)$, where D is a decision system.

$$R = \{X : X \subseteq C, \gamma_X(D) = \gamma_C(D)\} \tag{1.8}$$

$$R_{\min} = \{X : X \in R, \forall Y \in R, |X| \leq |Y|\} \tag{1.9}$$

The intersection of all of the reduced subsets is called the core, which contains all of the attributes that cannot be removed from the data set without introducing more contradictions.

$$Core(R) = \bigcap_{X \in R} X \tag{1.10}$$

Using the example shown in Table 1.2, the minimal reduct sets of C are as follows:

$$R = \{\{a,b\},\{b,c\}\{a,c\},\{a\},\{b\},\{c\}\}$$

The dependency degree of D = {d} on all of the possible reducts of C can be calculated as follows:

$$\gamma_C(D) = 1,$$

$$\gamma_{\{a,b\}}(D) = \frac{1}{6},\ \gamma_{\{b,c\}}(D) = 1,\ \gamma_{\{a,c\}}(D) = 1,$$

$$\gamma_{\{a\}}(D) = 0,\ \gamma_{\{b\}}(D) = \frac{1}{6},\ \gamma_{\{c\}}(D) = \frac{2}{3}$$

From these sets, the minimal reduct is as follows:

$$R_{\min} = \{\{a,b\},\{b,c\}\}$$

If the minimal reduct $\{b, c\}$ is selected, then the example data set presented in Table 1.2 can be reduced, as in Table 1.3.

.

Table1.3 Example of a reduced data set

U	***b***	***c***	***d***
1	0	0	1
2	0	1	0
3	0	2	0
4	0	0	1
5	1	1	1
6	0	2	0

It is obvious that determining all of the possible reducts is a time-consuming process, hence, this approach is practical for only small data sets. It is meaningless to calculate all of the reducts if the aim is to determine only one minimal reduct. Therefore, to improve the performance of the above method, an alternative strategy is required for large data sets.

Recently, many research efforts have focused on rough set theory as a reduction and classification method (Li & Cercone 2005; Shen & Chouchoulas 2000; Załuski et al. 2004). A review of the rough set theory approaches that have been applied to the attribute reduction problem can be found in (Thangavel & Pethalakshmi 2009).

1.7 THESIS OVERVIEW

This thesis is organised into seven chapters in accordance with the objectives mentioned above. This first chapter presents the background and motivation, the research objectives and scope. Chapter II presents a general review various search algorithms that have been applied to solve the attribute reduction problem.

Chapter III illustrates the research methodology which consists of five phases, namely, initial phase, preprocessing phase, construction phase, improvement phase and evaluation phase.

The initial phase concentrates on understanding the problem by reviewing the related work from the literature. The pre-processing phase concentrates on gathering the required data. In the construction phase, the initial solution is created for the different algorithms employed in this work. The quality of the solutions is further enhanced in the improvement phase by employing a number of meta-heuristic approaches. Lastly, in the evaluation phase, the performance of the proposed approaches is assessed by using the Wilcoxon test.

Chapter IV describes the employment of single-solution-based approaches to solve the attribute reduction problem. Four methods are proposed to solve this problem in two phases. In the first phase, the basic record-to-record travel algorithm and a modified Great Deluge algorithm are employed. Then, an enhancement is made to the proposed algorithms by incorporating a fuzzy logic controller in each algorithm in order to find the most suitable value for the parameter involved.

Chapter V presents a hybridisation of the local search algorithms with the GA, called the Fuzzy Memetic Algorithm. In this approach, the fuzzy record-to-record travel algorithm and fuzzy Great Deluge algorithm are hybridised with the GA. In order to make a significant comparison between the proposed approaches, a statistical analysis using the Wilcoxon test is undertaken.

Chapter VI compares proposed approaches and other available approaches in the literature. In addition, the classification accuracy and the number of generated rules based on the generated minimal attributes are also compared.

Finally, in Chapter VII, conclusions are drawn and the contributions of this research are set out. In addition, a number of areas to be pursued as future work are suggested. Samples of the solutions are given in the appendices at the end of this thesis.