A Survey on Evolutionary Computation Approaches to Feature Selection

Xue, Bing; Zhang, Mengjie; Browne, Will; Yao, Xin

DOI:
10.1109/TEVC.2015.2504420

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Xue, B, Zhang, M, Browne, W & Yao, X 2015, 'A Survey on Evolutionary Computation Approaches to Feature Selection', *IEEE Transactions on Evolutionary Computation*, no. 99. https://doi.org/10.1109/TEVC.2015.2504420

A Survey on Evolutionary Computation Approaches to Feature Selection

Bing Xue, *Member, IEEE,* Mengjie Zhang, *Senior Member, IEEE,* Will N. Browne, *Member, IEEE* and Xin Yao, *Fellow, IEEE*

***Abstract*—Feature selection is an important task in data mining and machine learning to reduce the dimensionality of the data and increase the performance of an algorithm, such as a classification algorithm. However, feature selection is a challenging task due mainly to the large search space. A variety of methods have been applied to solve feature selection problems, where evolutionary computation techniques have recently gained much attention and shown some success. However, there are no comprehensive guidelines on the strengths and weaknesses of alternative approaches. This leads to a disjointed and fragmented field with ultimately lost opportunities for improving performance and successful applications. This paper presents a comprehensive survey of the state-of-the-art work on evolutionary computation for feature selection, which identifies the contributions of these different algorithms. In addition, current issues and challenges are also discussed to identify promising areas for future research.**

***Index Terms*—Evolutionary computation, feature selection, classification, data mining, machine learning.**

I. Introduction

In data mining and machine learning, real-world problems often involve a large number of features. However, not all features are essential since many of them are redundant or even irrelevant, which may reduce the performance of an algorithm, e.g. a classification algorithm. Feature selection aims to solve this problem by selecting only a small subset of relevant features from the original large set of features. By removing irrelevant and redundant features, feature selection can reduce the dimensionality of the data, speed up the learning process, simplify the learnt model, and/or increase the performance [1], [2]. Feature construction (or feature extraction) [3], [4], [5], which can also reduce the dimensionality, is closely related to feature selection. The major difference is that feature selection selects a subset of original features while feature construction creates novel features from the original features. This paper focuses mainly on feature selection.

Feature selection is a difficult task due mainly to a large search space, where the total number of possible solutions is 2^n for a dataset with n features [1], [2]. The task is becoming more challenging as n is increasing in many areas with the advances in the data collection techniques and the increased complexity of those problems. An exhaustive search for the best feature subset of a given dataset is practically impossible in most situations. A variety of search techniques have been applied to feature selection, such as complete search, greedy search, heuristic search, and random search [1], [6], [7], [8], [9]. However, most existing feature selection methods still suffer from stagnation in local optima and/or high computational cost [10], [11]. Therefore, an efficient global search technique is needed to better solve feature selection problems. Evolutionary computation (EC) techniques have recently received much attention from the feature selection community as they are well-known for their global search ability/potential. However, there are no comprehensive guidelines on the strengths and weaknesses of alternative approaches along with their most suitable application areas. This leads to progress in the field being disjointed, shared best practice becoming fragmented and ultimately, opportunities for improving performance and successful applications being missed. This paper presents a comprehensive survey of the literature on EC for feature selection with the goal of providing interested researchers with the state-of-the-art research.

Feature selection has been used to improve the quality of the feature set in many machine learning tasks, such as classification, clustering, regression, and time series prediction [1]. This paper focuses mainly on feature selection for classification since there is much more work on feature selection for classification than for other tasks [1]. Recent reviews on feature selection can be seen from [7], [8], [12], [13], which focus mainly on non-EC based methods. De La Iglesia [14] presents a summary of works using EC for feature selection in classification, which is suitable for a non-EC audience since it focuses on basic EC concepts and genetic algorithms (GAs) for feature selection. The paper [14] reviewed only 14 papers published since 2010 and in total 21 papers since 2007. No papers published in the most recent two years were reviewed [14], but there have been over 500 papers published in the last five years. Research on EC for feature selection started around 1990, but it has become popular since 2007 when the number of features in many areas became relatively large. Fig. 1 shows the number of papers on the two most popular EC methods in feature selection, i.e. GAs and particle swarm optimisation (PSO), which shows that the number of papers, especially on PSO, has significantly increased since 2007 (Note that the numbers were obtained from Google Scholar on September 2015. These numbers might not be complete, but they show the general trend of the field. The papers used to form this survey were collected from all the major databases, such as Web of Science, Scopus, and Google

Bing Xue, Mengjie Zhang, and Will N. Browne are with the Evolutionary Computation Research Group at Victoria University of Wellington, PO Box 600, Wellington, New Zealand (E-mail: bing.xue@ecs.vuw.ac.nz).
Xin Yao is with the Natural Computation Group, School of Computer Science at The University of Birmingham, Edgbaston, Birmingham B15 2TT, U.K.

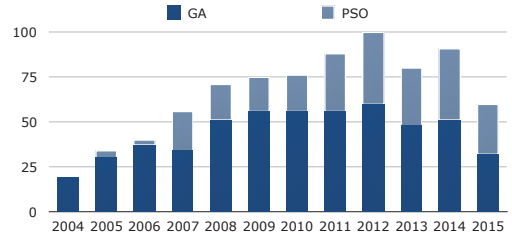

Fig. 1. Number of Papers on GAs and PSO for Feature Selection (from Google Scholar, September 2015).

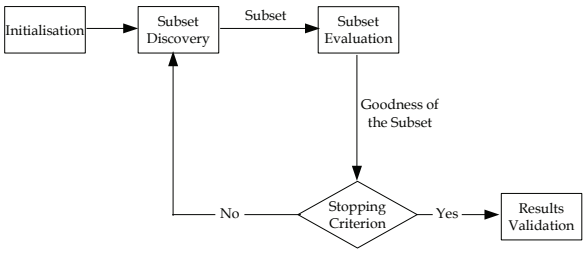

Fig. 2. General Feature Selection Process [1].

Scholar). We aim to provide a comprehensive survey of the state-of-the-art work and a discussion of the open issues and challenges for future work. We expect this survey to attract attention from researchers working on different EC paradigms to further investigate effective and efficient approaches to addressing new challenges in feature selection. This paper is also expected to encourage researchers from the machine learning community, especially classification, to pay much attention to the use of EC techniques to address feature selection problems.

The remainder of this paper is organised as follows. Section II describes the background of feature selection. Section III reviews typical EC algorithms for feature selection. Section IV discusses different measures used in EC for feature selection. Section V presents the applications of EC based feature selection approaches. Section VI discusses current issues and challenges, and conclusions are given in Section VII.

II. Background

Feature selection is a process that selects a subset of relevant features from the original large set of features [9]. For example, feature selection is to find key genes (i.e. biomarkers) from a large number of candidate genes in biological and biomedical problems [15], to discover core indicators (features) to describe the dynamic business environment [9], to select key terms (features, e.g. words or phrases) in text mining [16], and to choose/construct important visual contents (features, e.g. pixel, color, texture, shape) in image analysis [17]. Fig. 2 shows a general feature selection process and all the five key steps, where "Subset Evaluation" is achieved by using an evaluation function to measure the goodness/quality of the selected features. Detailed discussions about Fig. 2 can be seen in [1] and a typical iterative evolutionary workflow of feature selection can be seen in [18].

Based on the evaluation criteria, feature selection algorithms are generally classified into two categories: *filter* approaches and *wrapper* approaches [1], [2]. Their main difference is that wrapper approaches include a classification/learning algorithm in the feature subset evaluation step. The classification algorithm is used as a "black box" by a wrapper to evaluate the goodness (i.e. the classification performance) of the selected features. A filter feature selection process is independent of any classification algorithm. Filter algorithms are often computationally less expensive and more general than wrapper algorithms. However, filters ignore the performance of the selected features on a classification algorithm while wrappers evaluate the feature subsets based on the classification performance, which usually results in better performance achieved by wrappers than filters for a particular classification algorithm [1], [7], [8]. Note that some researchers categorise feature selection methods into three groups: wrapper, embedded and filter approaches [7], [8]. The methods integrating feature selection and classifier learning into a single process are called embedded approaches. Among current EC techniques, only genetic programming (GP) and learning classifier systems (LCSs) are able to perform embedded feature selection [19], [20]. Thus, to simplify the structure of the paper, we follow the convention of classifying feature selection algorithms into wrappers and filters only [1], [2], [21] with embedded algorithms belonging to the wrapper category.

Feature selection is a difficult problem not only because of the large search space, but also feature interaction problems. Feature interaction (or epistasis [22]) happens frequently in many areas [2]. There can be two-way, three-way or complex multi-way interactions among features. A feature, which is weakly relevant to the target concept by itself, could significantly improve the classification accuracy if it is used together with some complementary features. In contrast, an individually relevant feature may become redundant when used together with other features. The removal or selection of such features may miss the optimal feature subset(s). Many traditional measures evaluating features individually cannot work well and a subset of features needs to be evaluated as a whole. Therefore, the two key factors in a feature selection approach are the *search techniques*, which explore the search space to find the optimal feature subset(s), and the *evaluation criteria*, which measure the quality of feature subsets to guide the search.

Feature selection involves two main objectives, which are to maximise the classification accuracy and minimise the number of features. They are often *conflicting objectives*. Therefore, feature selection can be treated as a multi-objective problem to find a set of trade-off solutions between these two objectives. The research on this direction has gained much attention only in recent years, where EC techniques contribute the most since EC techniques using a population based approach are particularly suitable for multi-objective optimisation.

A. Existing Work on Feature Selection

This section briefly summarises them from three aspects, which are the search techniques, the evaluation criteria, and the number of objectives.

1) Search techniques: There are very few feature selection methods that use an exhaustive search [1], [7], [8]. This is because even when the number of features is relatively small

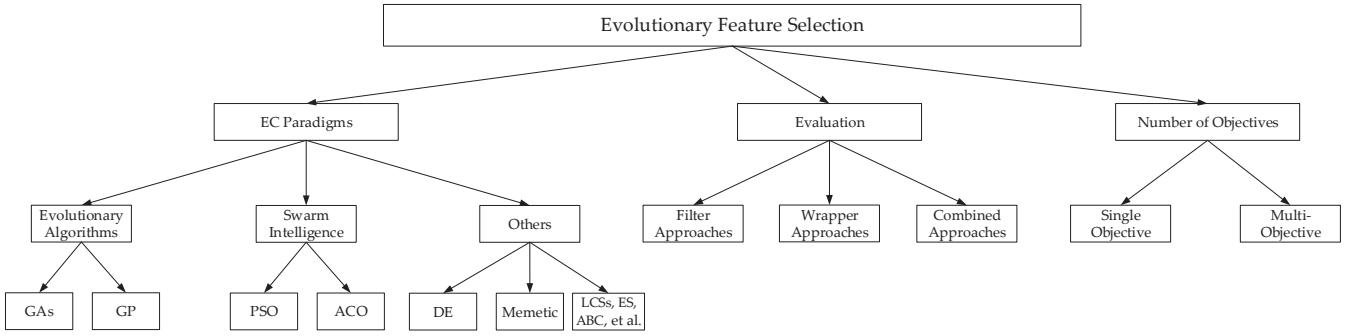

Fig. 3. Overall categories of EC for feature selection.

(e.g. 50), in many situations such methods are computationally too expensive to perform. Therefore, different heuristic search techniques have been applied to feature selection, such as greedy search algorithms, where typical examples are sequential forward selection (SFS) [23], sequential backward selection (SBS) [24]. However, both methods suffer from the so-called "nesting effect" because a feature that is selected or removed cannot be removed or selected in later stages. "plus-l-take-away-r" [25] compromises these two approaches by applying SFS l times and then SBS r times. This strategy can avoid the nesting effect in principle, but it is hard to determine appropriate values for l and r in practice. To avoid this problem, two methods called sequential backward floating selection (SBFS) and sequential forward floating selection (SFFS) were proposed in [26]. Both floating search methods are claimed to be better than the static sequential methods. Recently, Mao and Tsang [27] proposed a two-layer cutting plane algorithm to search for the optimal feature subsets. Min et al. [28] proposed a heuristic search and a backtracking algorithm, which performs exhaustive search, to solve feature selection problems using rough set theory. The results show that heuristic search techniques achieved similar performance to the backtracking algorithm, but used a much shorter time. In recent years, EC technique as they are effective methods have been applied to solve feature selection problems. Such methods include GAs, GP, particle swarm optimisation (PSO), and ant colony optimisation (ACO). Details will be described in the next section.

Feature selection problems have a large search space, which is often very complex due to feature interaction. Feature interaction leads to individually relevant features becoming redundant or individually weakly relevant features becoming highly relevant when combined with other features. Compared with traditional search methods, EC techniques do not need domain knowledge and do not make any assumptions about the search space, such as whether it is linearly or non-linearly separable, and differentiable. Another significant advantage of EC techniques is that their population based mechanism can produce multiple solutions in a single run. This is particularly suitable for multi-objective feature selection in order to find a set of non-dominated solutions with the trade-off between the number of features and the classification performance. However, EC techniques have a major limitation of requiring a high computational cost since they usually involve a large number of evaluations. Another issue with EC techniques is their stability since the algorithms often select different features from different runs, which may require a further selection process for real-world users. Further research to address these issues is of great importance as the increasingly large number of features increases the computational cost and lowers the stability of the algorithms in many real-world tasks.

2) Evaluation criteria: For wrapper feature selection approaches, the classification performance of the selected features is used as the evaluation criterion. Most of the popular classification algorithms, such as decision tree (DT), support vector machines (SVMs), Naïve Bayes (NB), K-nearest neighbour (KNN), artificial neural networks (ANNs), and linear discriminant analysis (LDA), have been applied to wrappers for feature selection [7], [8], [29]. For filter approaches, measures from different disciplines have been applied, including information theory based measures, correlation measures, distance measures, and consistency measures, and [1].

Single feature ranking based on a certain criterion is a simple filter approach, where feature selection is achieved by choosing only the top-ranked features [7]. Relief [30] is a typical example, where a distance measure is used to measure the relevance of each feature and all the relevant features are selected. Single feature ranking methods are computationally cheap, but do not consider feature interactions, which often leads to redundant feature subsets (or local optima) when applied to complex problems, e.g. microarray gene data, where genes possess intrinsic linkages [1], [2]. To overcome such issues, filter measures that can evaluate the feature subset as a whole have become popular. Recently, Wang et al. [31] developed a distance measure evaluating the difference between the selected feature space and all feature space to find a feature subset, which approximates all features. Peng et al. [32] proposed the minimum Redundancy Maximum Relevance method based on mutual information, where the proposed measures have been introduced to EC for feature selection due to their powerful search abilities [33], [34].

Mao and Tsang [27] proposed a novel feature selection approach by optimizing multivariate performance measures (which can also be viewed as an embedded method since the proposed feature selection framework was to optimise the general loss function and was achieved based on SVMs). However, the proposed method resulted a huge search space for high-dimensional data, which required a powerful heuristic search method to find the optimal solutions. Statistical approaches, such as T-test, logistic regression, hierarchical clustering, and

CART, are relatively simple and can achieve good performance [35]. Sparse approaches have recently become popular, such as sparse logistic regression for feature selection [36], which has been used for feature selection tasks with millions of features. For example, the sparse logistic regression method [36] automatically assigns a weight to each feature showing its relevance. Irrelevant features are assigned with low weights close to zero, which has the effect of filtering out these features. Sparse learning based methods tend to learn simple models due to their bias to features with high weights. These statistical algorithms usually produce good performance with high efficiency, but they often have assumptions about the probability distribution of the data. Furthermore, the used cutting plan search method in [36] works well when the search space is unimodal, but EC approaches can deal well with both unimodal and multimodal search space and the population based search can find a Pareto front of non-dominated (trade-off) solutions. Min et al. [28] developed a rough set theory based algorithm to address feature selection problems under the constraints of having limited resources (e.g. money and time). However, many studies show that filter methods do not scale well to problems with more than tens of thousands of features [13].

3) Number of objectives: Most of the existing feature selection methods aim to maximise the classification performance only during the search process or aggregate the classification performance and the number of features into a single objective function. To the best of our knowledge, all the multi-objective feature selection algorithms to date are based on EC techniques since their population based mechanism producing multiple solutions in a single run is particularly suitable for multi-objective optimisation.

B. Detailed Coverage of This Paper

As shown in Fig. 3, according to three different criteria, which are the *EC paradigms*, the *evaluation*, and the number of *objectives*, EC based feature selection approaches are classified into different categories. These three criteria are the key components in a feature selection method. EC approaches are mainly used as the search techniques in feature selection. Almost all the major EC paradigms have been applied to feature selection and the most popular ones are discussed in this paper, i.e. GAs [37], [38], [39] and GP [19], [40], [41] as typical examples in evolutionary algorithms, PSO [10], [29], [42] and ACO [43], [44], [45], [46] as typical examples in swarm intelligence, and other algorithms recently applied to feature selection, including differential evolution (DE) [47], [48][1], memetic algorithms [49], [50], LCSs [51], [52], evolutionary strategy (ES) [53], artificial bee colony (ABC) [54], [55], and artificial immune systems (AISs) [56], [57]. Based on the evaluation criteria, we review both filter and wrapper approaches, and also include another group of approaches named "Combined". "Combined" means that the evaluation procedure includes both filter and wrapper measures, which are also called *hybrid* approaches by some researchers [9], [14]. The use here of "Combined" instead of "hybrid" is to avoid confusion with the concept of *hybrid* algorithms in the EC field, which hybridise multiple EC search techniques. According to the number of objectives, EC based feature selection approaches are classified into single objective and multi-objective approaches, where the multi-objective approaches correspond to methods aiming to find a Pareto front of trade-off solutions. The approaches that aggregate the number of features and the classification performance into a single fitness function are treated as single objective algorithms in this paper.

[1]Some researchers classify DE as a swarm intelligence algorithm.

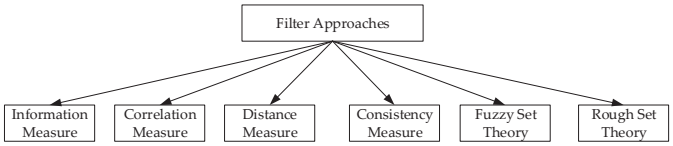

Fig. 4. Different measures in EC based filter approaches.

Similar to many earlier survey papers on traditional (non-EC) feature selection [1], [7], [8], [9], this paper further reviews different evolutionary filter methods according to measures that are driven from different disciplines. Fig. 4 shows the main categories of measures used in EC based filter approaches. Wrapper approaches are not further categorised according to their measures because the classification algorithm in wrappers is used as a "black box" during the feature selection process such that it can often be easily replaced by another classification algorithm.

The reviewed literature is organised as follows. Typical approaches are reviewed in Section III, where each subsection discusses a particular EC technique for feature selection (e.g. Section III-A: GAs for feature selection, as shown by the left branch in Fig. 3). Within each subsection, the research using an EC technique is further detailed and discussed according to the evaluation criterion and the number of objectives. In addition, Section IV discusses the research on EC based filter approaches for feature selection. The applications of EC for feature selection are described in Section V.

TABLE I
CATEGORISATION OF GA APPROACHES

	Single Objective	Multi-Objective
Wrapper	[3], [37], [58], [38], [39], [44], [59], [60], [61], [62], [63], [64], [65], [66], [67], [68], [69], [70], [71], [72], [73], [74], [75], [76], [77], [78], [79], [80], [81], [82], [83], [84], [85], [86], [87]	[88], [89], [90], [91], [92], [93], [94], [95], [96], [97]
Filter	[75], [98], [99], [100], [101], [102]	[102], [103], [104], [105], [106]
Combined	[107], [108], [109]	

III. EC FOR FEATURE SELECTION

A. GAs for Feature Selection

GAs are most likely the first EC technique widely applied to feature selection problems. One of the earliest works was published in 1989 [37]. GAs have a natural representation of a binary string, where "1" shows the corresponding feature is selected and "0" means not selected. Table I shows the typical works on GAs for feature selection. It can be seen that there are more works on wrappers than filters, and more on single objective than multi-objective approaches.

For wrapper approaches, different classification algorithms have been used to evaluate the goodness of the selected features, e.g. SVMs [68], [71], [72], [73], [75], [79], [80], [81], [86], [107], KNN [39], [74], [76], [77], [80], [81], [86], [95], [107], ANNs [61], [69], [78], [81], [83], [85], DT [60], [80], [107], NB [80], [107], [109], multiple linear regression for classification [59], extreme learning machines (ELMs) [110], and discriminant analysis [66], [67], [82]. SVMs and KNN are the most popular classification algorithms due to their promising classification performance and simplicity, respectively. For filter approaches, different measures have been applied to GAs for feature selection, e.g. information theory [102], [105], [106], consistency measures [98], [105], rough set theory [103] and fuzzy set theory [99].

Many different new enhancements to GAs have been proposed to improve the performance, which focus mainly on the search mechanisms, the representation, and the fitness function. Some early works [59], [62] introduced GAs to feature selection by investigating the influence of the population size, mutation, crossover, and reproduction operators, but with limited experiments.

Recently, Derrac et al. [76] proposed a cooperative coevolutionary algorithm for feature selection based on a GA with three populations, where the first focused on feature selection, the second focused on instance selection, and the third focused on both feature selection and instance selection. The proposed algorithm addressed feature selection and instance selection in a single process, which reduced the computational time. Such approaches should be further investigated in the future given that large datasets (i.e. with thousands or tens of thousands of features) may include not only irrelevant features, but also noisy instances. Li et al. [75] also proposed a multiple populations based GA for feature selection, where every two neighbour populations shared two individuals to exchange information for increasing the search ability. Local search was performed on the best individual in each population to further increase the performance. The proposed scheme was tested with different filter and wrapper measures, and was shown to be effective for feature selection, but it was tested only on datasets with a maximum number of 60 features.

Chen et al. [71] proposed to address feature selection problems through GAs for feature clustering, where a GA was used to optimise the cluster centre values of a clustering method to group features into different clusters. Features in each cluster were then ranked according to their distance values to the cluster centre. Feature selection was achieved by choosing the top-ranked features as representatives from each cluster. The proposed algorithm was shown to be effective on datasets with thousands of features. Lin et al. [111] proposed a GA based feature selection algorithm adopting domain knowledge of financial distress prediction, where features were classified into different groups and a GA was used to search for feature subsets consisting of top candidate features from each group. This work may have a similar problem to [71] in terms of ignoring feature interactions. A GA was used in a two-stage approach, where a filter measure was used to rank features and only the top-ranked ones were used in GA based feature selection [72], [85]. In contrast, Zamalloa et al. [67] used a GA to rank features directly and feature selection was achieved by choosing only the top-ranked features. A potential limitation in [67], [72] and [85] is that the removed lowly-ranked features might become highly useful when combined with other features because of feature interaction.

Traditional feature selection methods have also been adopted in GAs to improve the performance. Jeong et al. [39] developed a partial SFFS mutation operator in GAs, where SFFS was performed to improve the feature subset selected by a chromosome. The proposed algorithm was shown to be effective for feature selection, but it may have a potential problem of being computationally expensive due to the extra calculation of SFFS. Gheyas and Smith [38] developed a hybrid algorithm named SAGA based on a GA and simulated annealing (SA), and compared it with different EC algorithms and traditional methods for feature selection, including a GA, ACO, PSO, SFS, SFFS, SFBS, and SA. The results showed that SAGA performed the best in terms of the classification performance. The combination of the global search ability of GAs and the local search ability of SA may be the reason for the superior performance of SAGA.

In terms of representation, Hong and Cho [69] proposed a binary vector to represent each chromosome (i.e. genotype), where a predefined small number (pd) of binary bits are converted to an integer number i indicating that the ith feature is selected. Therefore, the length of the representation/genotype was determined by multiplying pd and the desired number of features. It reduced the dimensionality of the GA search space on high-dimensional datasets with thousands of features, which resulted in better performance than the traditional representation. Chen et al. [82] also developed a binary representation, which included two parts, where the first part was converted to an integer representing the number of features to be selected while the second showed which features were selected. Jeong et al. [39] proposed a new representation to further reduce the dimensionality, where the length of the chromosome was equal to the number of desired features. The values in chromosomes indicated the indexes of features. When the index of a feature(s) appeared multiple times, a partial SFFS operator was applied to choose alternative features to avoid duplication. One limitation in [39], [69] and [82] is that the number of features needs to be pre-defined, which might not be the optimal size. To address this limitation, Yahya et al. [112] developed a variable length representation, where each chromosome showed the selected features only and different chromosomes may have different lengths. New genetic operators were accordingly developed to cope with the variable length representation. However, the performance of the proposed algorithm was not compared with other GAs based methods.

Li et al. [77] proposed a bio-encoding scheme in a GA, where each chromosome included a pair of strings. The first string was binary-encoded to show the selection of features and the second was encoded as real-numbers indicating the weights of features. By combining this with an Adaboost learning algorithm, the bio-encoding scheme achieved better performance than the traditional binary encoding. Winkler et al. [81] proposed a new representation that included both feature

selection and parameter optimisation of a certain classification algorithm, e.g. an SVM. The length was the total number of features and parameters. Souza et al. [83] developed a three-level representation in a GA and multilayer perceptron (MLP) for feature selection, which indicated the selection of features, the pruning of the neurons, and the architecture of the MLP, respectively. These three examples [77], [81], [83] suggest that combining the selection of features and the optimisation of a classification algorithm is an effective way to improve the classification performance since both the data and the classifier are optimised, which can also be evident from [66], [71], [79].

In terms of the fitness function, Da Silva et al. [80] aggregated the classification accuracy and the number of features into a single fitness function. Yang and Honavar [61] proposed to combine the maximisation of the classification accuracy and the minimisation of the cost of an ANN into a single fitness function. Winkler et al. [81] proposed several fitness functions, which considered the number of features, the overall classification performance, the class specific accuracy, and the classification accuracy using all the original features. Sousa et al. [109] employed a fitness function using area under curve (AUC) of the receiver operating characteristic (ROC) of a NB classifier. In [107], a filter measure (Pearson correlation measure) and a wrapper measure (classification accuracy) were combined to form a single fitness function in a GA for feature selection to utilise the advantages of each measure.

GAs for multi-objective feature selection started much later (around 10 years later) than for single objective feature selection. Most of the multi-objective approaches are based on non-dominated sorting GA II (NSGA-II) or its variations [92], [94], [96], [97], [102], [103], [105]. Although there are more works on multi-objective feature selection using GAs than using other EC techniques, the potential of GAs for multi-objective feature selection has still not been thoroughly investigated since feature selection is a complex task that requires specifically designed multi-objective GAs to search for the non-dominated solutions.

In summary, GAs have been applied to feature selection for around 25 years and have achieved reasonably good performance on problems with hundreds of features. Researchers introduced GAs to address feature selection problems including thousands of features with limited success, where most are wrapper approaches. This leads to a high computational cost since GAs usually involve a large number of evaluations and each evaluation in a wrapper approach usually takes a relatively long time, especially when the number of instances is large. As a result, although GAs approaches have been proposed for some feature selection tasks with thousands of features, almost all these feature selection tasks have a relatively small number of instances, i.e. less than one thousand [70], [71]. Such approaches struggle to solve "big data" tasks, where both the number of features and the number of instances are huge. This is not only an issue for GAs, but also for other EC techniques for feature selection. To use GAs to address such tasks, a novel representation which can reduce the dimensionality of the search space will be needed. The design of genetic operators, e.g. crossover and mutation, provides opportunities to identify good building blocks (i.e. feature groups) and combine or adjust complementary features to find optimal feature subsets, but this is a challenging task. Furthermore, when and how to apply these operators, and the parameter settings in GAs are also key factors influencing their performance on feature selection.

TABLE II
CATEGORISATION OF GP APPROACHES

	Single Objective	Multi-Objective
Wrapper	[19], [20], [40], [113], [114], [115], [116], [117], [15], [118], [119], [120], [121], [122], [123], [124], [125], [126]	
Filter	[41], [127], [128]	[129], [130]

B. GP for Feature Selection

Table II shows typical works on GP for feature selection. Compared with GAs and PSO, there are a much smaller number of works on GP for feature selection. GP is used more often in feature construction than feature selection because of its flexible representation. In feature selection, most GP works use a tree-based representation, where the features used as the leaf nodes of a tree are the selected features. GP can be used as a search algorithm and also a classification algorithm. In filter approaches, GP is mainly used as the search algorithm. In most wrapper (or embedded) approaches, GP is used as both the search method and the classification algorithm. In a very few cases, GP was used as a classification algorithm only in a feature selection approach [81].

One of the early works on GP for feature selection was published in 1996 [123], where a generalised linear machine was used as the classifier to evaluate the fitness of the selected features. Later, Neshatian and Zhang [128] proposed a wrapper feature selection approach based on GP, where a variation of NB algorithm was used for classification. A bit-mask encoding was used to represent feature subsets. Set operators were used as primitive functions. GP was used to combine feature subsets and set operators together to find an optimal subset of features. Hunt et al. [115] developed GP based hyper-heuristics for wrapper feature selection, where two function operators for removing and adding features were proposed. The results showed that the proposed algorithm improved the classification performance and reduced the number of features over using standard GP.

For filter approaches, GP was applied to feature selection using an improved information theory based measure [41]. Further, a GP based relevance measure was proposed to evaluate and rank feature subsets [130], which was a single objective algorithm but could provide solutions with different sizes and accuracies. However, it may suffer from the problem of high computational cost.

In most works, GP was used to search for the optimal feature subset and simultaneously trained as a classifier. Muni et al. [19] developed a wrapper feature selection model based on multi-tree GP, which simultaneously selected a good feature subset and learned a classifier using the selected features. Two new crossover operations were introduced to increase the performance of GP for feature selection. Based on the two crossover operations introduced by Muni et al. [19], Purohit

et al. [40] further introduced another crossover operator, which was randomly performed for selecting a subtree from the first parent and finding its best place in the second parent. Both [19] and [40] showed the powerful ability of GP for simultaneously performing feature selection and learning a classifier. They are similar to the works [71], [79] using GAs to simultaneously select features and optimise a classifier, but the main difference is that GP itself was used as both a search technique for selecting features and as a classifier for classification (i.e. embedded feature selection) while GAs were used as search techniques only.

Two-stage approaches have been investigated in GP for feature selection. Venkatraman et al. [124] proposed to use a mutual information measure to rank individual features and remove weakly relevant or irrelevant features in the first stage and GP was then applied to select a subset of the remaining features [124]. Later, to take different advantages of different measures, multiple filter measures were used to rank features and a set of features were selected according to each measure. The combination of these features was used as input to GP for further feature selection [116], [119]. However, a potential limitation is that individual feature ranking may remove potentially useful features without considering their interactions with other features. Neshatian and Zhang [120] proposed another type of individual feature ranking approach, where each feature was given a score according to its frequency of appearance in the best GP individuals. Feature selection was achieved by using only the top-ranked features for classification. This way of evaluating individual features took other features into account, which could avoid the limitation of most single feature ranking methods.

A GP based multi-objective filter feature selection approach was proposed for binary classification problems [129]. Unlike most filter methods that usually could measure only the relevance of a single feature to the class labels, the proposed algorithm could discover hidden relationships between subsets of features and the target classes, and achieve better classification performance. There are only a few works on GP for multi-objective feature selection. It will be interesting to investigate this in the future since GP has shown its ability in addressing feature selection and multi-objective problems [131].

In summary, GP for feature selection has achieved some success, but with much less work than GAs. Compared with GAs, GP is able to perform embedded feature selection to be used as both a search technique and a classifier. GAs are easier to implement and have a straightforward representation easily indicating the selection of features, which can be a good choice for relatively low-dimensional problems, e.g. less than one thousand. Due to the flexible representation, GP can also perform feature construction to create new high-level features to further increase the classification performance, and GP has a potential to handle large-scale feature selection since the representation does not have to include the selection information (or the index) of all features. Further, many real-world problems, such as gene selection, include a large number (i.e. tens of thousands) of features, but a very small number (less than 100) of instances, which is a challenge not only in machine learning, but also in statistics and biology. GP can handle tasks with a very small number of instances [132], which provides an opportunity to better solve feature selection tasks with a small number of instances. When and how to apply genetic operators is also important in GP, but the design and the use of the genetic operators in GP is more difficult than in GAs due to the flexible representation and the different return types of the functions. The parameter settings in GP is also very important. Because of the large population size, GP may suffer from the issue of being computationally expensive.

TABLE III
CATEGORISATION OF PSO APPROACHES

	Single Objective	Multi-Objective
Wrapper	[10], [42], [70], [133], [134], [135], [136], [137], [138] [139], [140], [141], [142], [143], [144], [145], [146], [147], [148], [149], [150], [151], [152], [153], [154], [155], [156], [157], [158], [159], [160]	[29], [161], [162]
Filter	[34], [163], [164], [165], [166], [167], [168], [169], [170]	[171], [172], [173], [174]
Combined	[11], [33], [175], [176], [177]	

C. PSO for Feature Selection

Both continuous PSO and binary PSO have been used for both filter and wrapper, single objective and multi-objective feature selection. The representation of each particle in PSO for feature selection is typically a bit-string, where the dimensionality is equal to the total number of features in the dataset. The bit-string can be binary numbers in binary PSO or real-value numbers in continuous PSO. When using binary representation, "1" means the corresponding feature is selected and "0" means not. When using the continuous representation, a threshold θ is usually used to determine the selection of a particular feature, i.e. if the value is larger than θ, the corresponding feature is selected. Otherwise, it is not selected.

As can be seen from Table III, there has been more research on PSO for single objective than multi-objective, and more research on wrapper than filter feature selection. For wrapper approaches, different classification algorithms have been used with PSO to evaluate the goodness of the selected features, e.g. SVMs [33], [70], [134], [135], [137], [141], KNN [148], [149], [151], [152], [160], LDA [144], ANNs [42], [147], [178], logistic regression classification model [10], and Adaboost [142]. SVMs and KNN are the most popular classification algorithms because of their promising classification performance and simplicity, respectively. For filter approaches, different measures have been applied to PSO for feature selection and details can be seen in Section IV.

A number of new PSO algorithms have been proposed to improve performance on feature selection problems, including initialisation strategies, representation, fitness functions, and the search mechanisms. Xue et al. [158] developed a new initialisation strategy to mimic the typical forward and backward feature selection methods in the PSO search process, which showed that good initialisation significantly increased the performance of PSO for feature selection.

There are only a few works on developing new representations in PSO for feature selection. The typical representation has been slightly modified to simultaneously perform feature

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TEVC.2015.2504420

selection and parameter optimsation of a classification algorithm, mostly optimising the parameters in the kernel functions of SVMs [138], [141], [153], [179]. The length of the new representation is equal to the total number of features and parameters. The representation was encoded in three different ways, being continuous encoding [138], binary encoding [153], and a mixture of binary and continuous encoding [141], [179]. Since PSO was originally proposed for continuous optimisation, continuous encoding performed better than the other two encoding schemes. Lane et al. [154] proposed the use of PSO and statistical clustering (which groups similar features into the same cluster) for feature selection, where a new representation was proposed to incorporate statistical feature clustering information during the search process of PSO. In the new representation, features from the same cluster were arranged together and only a single feature was selected from each cluster. The proposed algorithm was shown to be able to significantly reduce the number of features. Lane et al. [157] further improved the algorithm by allowing the selection of multiple features from the same cluster to further improve the classification performance. Later, Nguyen et al. [155] proposed a new representation, where the dimensionality of each particle was determined by the maximum number of desired features. The dimensionality of the new representation is much smaller than the typical representation, however, it is not easy to determine the desired number of features.

Learning from neighbours' experience, i.e. social interaction through $gbest$, and learning from each individual's own experience through $pbest$, are the key ideas in PSO. Chuang et al. [140] developed a $gbest$ resetting mechanism by including zero features in order to guide the swarm to search for small feature subsets. Xue et al. [158] considered the number of features when updating $pbest$ and $gbest$ during the search process of PSO, which could further reduce the number of features over the traditional updating $pbest$ and $gbest$ mechanism without deteriorating the classification performance. Tran et al. [156] used the $gbest$ resetting mechanism in [140] to reduce the number of features and performed a local search process on $pbest$ to increase the classification performance. Each evaluation in the local search was sped up by calculating fitness based only on the features being changed (from selected to not selected or from not selected to selected) instead of based on all the selected features. The proposed algorithm further reduced the number of features and improved the classification performance over the previous method [140] and standard PSO. PSO with multiple swarms to share experience has also been applied to feature selection [11], [180], but may lead to the problem of high computational cost.

The fitness function plays an important role in PSO for feature selection. For filter approaches, the fitness function is formed by using different measures, which will be discussed in detail in Section IV. For wrapper approaches, many existing works used only the classification performance as the fitness function [11], [134], [135], [137], [138], [139], [140], [142], [160], which led to relatively large feature subsets. However, most of the fitness functions used different ways to combine both the classification performance and the number of features into a single fitness function [70], [136], [141], [180], [148], [181]. However, it is difficult to determine in advance the optimal balance between them without *a priori* knowledge. Multi-objective feature selection can help solve this problem by simultaneously optimising these two objectives to obtain a set of trade-off solutions.

Research on PSO for multi-objective feature selection started only in the last two years, where Xue et al. [29], [161] conducted the first work to optimise the classification performance and the number of features as two separate objectives. Continuous and binary PSO in multi-objective feature selection were directly compared in [161], where the results showed that continuous PSO can usually achieve better performance than binary PSO since binary PSO has potential limitations, such as the position of a particle in binary PSO is updated solely based on the velocity while the position in standard PSO is updated based on both the velocity and current position [182]. Further, the performance of the multi-objective PSO algorithm for feature selection was compared with three other popular evolutionary multi-objective algorithms, NSGAII, SPEA2 and PAES [29], where the multi-objective PSO approach was shown to be superior to the other three methods. PSO was also applied to multi-objective filter feature selection, where information based theory [171], [173] and rough set theory [172], [174] were used to evaluate the relevance of the selected features. These works showed that PSO for multi-objective feature selection provided multiple and better solutions/choices to users.

To sum up, there has been rapid development on PSO for feature selection. PSO has a similar advantage to GAs in terms of a straightforward representation, but neither of them can be used for feature construction (unlike GP with its flexible representation). However, the representation of GAs and PSO might not scale well on problems with thousands or tens of thousands of features, since it forms a huge search space. In other important aspects GAs and PSO take different approaches to evolving good feature subsets. GAs address combinatorial optimisation problems by identifying good building blocks of information, combining complementary blocks via crossover and adjustment via mutation. Thus, GAs are likely to be suited to domains where there are groups of interacting features, potentially with multiple good subsets, to consider. PSO has a more structured neighbourhood guiding its recombination method than GAs, as well as a velocity term that enables fast convergence to a solution. PSO should suit domains where there is a structure in how features interact, i.e. low sensitivity to the inclusion of each feature in a solution, and where fast convergence does not lead to local optima. PSO has an advantage over GAs and GP of being easy to implement. Developing novel PSO algorithms, particularly novel search mechanisms, parameter control strategies and representation, for large-scale feature selection, is still an open issue.

D. ACO for Feature Selection

Table IV shows typical works on ACO for feature selection, where the earliest work was proposed around 2003 [183]. Table IV shows that there are more papers on wrapper methods than filter and embedded methods. Most of work focuses on

TABLE IV
CATEGORISATION OF ACO APPROACHES

	Single Objective	Multi-Objective
Wrapper	[43], [44], [45], [46], [184], [185], [186], [187], [188], [16], [189], [190], [191], [192], [193]	[194]
Filter	[183], [195], [196], [197], [198], [199], [200], [201], [202], [203]	[204]
Combined	[47], [205]	

single objective methods and there are only a few papers on multi-objective approaches.

In one of the early works, ACO and an SVM were used for wrapper feature selection for face recognition, where the original features were extracted by principal component analysis (PCA) from the images in the preprocessing stage [43]. Ke et al. [197] proposed the use of limited pheromone values in ACO for feature selection and the proposed algorithm also updated the pheromone trails of the edges connecting every two different features of the best-so-far solution. Experimental results showed that the proposed algorithms achieved better performance than SA, a GA, and Tabu search based algorithms in terms of both the classification performance and the number of features. O'Boyle et al. [188] proposed to use ACO to simultaneously select features and optimise the parameters of an SVM, where a weighting method was also proposed to determine the probability of an ant selecting a particular feature. Khushaba et al. [47] combined ACO and DE for feature selection, where DE was used to search for the optimal feature subset based on the solutions obtained by ACO. A traditional feature selection algorithm, forward selection, was also introduced to ACO [206], where ACO started with a small set of core features. Vieira et al. [190] proposed a cooperative ACO algorithm with two colonies for feature selection, where the first one decided the number of features needed and the second colony was to select individual features. Santana et al. [44] compared the performance of ACO with a GA based feature selection method for ensemble classifiers. The results showed that ACO performed better when the number of individual classifiers was small while the GA performed better when this number was large.

The representation of ACO for feature selection is typically a graph, where features are encoded as nodes to construct a graph model. Each ant represents a feature subset, where the features selected are the nodes it visited. In most ACO based algrithms [188], [16], features/nodes are fully connected to each other in the graph, but in [189], each feature was connected only to two features. The final solution [189] was a binary set, whose length was equal to the number of nodes (features) that the ant visited. The value of "1" means the corresponding feature is selected and "0", otherwise. Chen et al. [45] proposed a new representation scheme to reduce the size of the search space (i.e. graph), where each feature/node was connected only to the next node using two edges showing "selected" or "not selected". This representation scheme significantly reduced the total number of edges that ACO needed to traverse. Kashef and Nezamabadi [192], [193] also proposed a new representation, where each feature had two nodes, one for selecting that feature and the other for removing. At the end of a tour, each ant had a binary vector with the length as the total number of features, where "1" indicated selecting and "0" indicated removing the corresponding feature.

In most ACO based wrapper approaches, the classification performance was used as the fitness evaluation criterion. In [185], [47], the fitness of ants (feature subsets) was evaluated using the overall classification performance, but the performance of individual features was also considered to further improve the performance. The fitness functions in [187], [16] included both the classification performance and the number of features. Later, by extending the work on single objective ACO and a fuzzy classifier for feature selection [186], Vieira et al. [194] developed a multi-objective wrapper approach, where ACO aimed to minimise both the classification error and the number of features. Recently, Ke et al. [204] developed new multi-objective ACO for filter feature selection, which adopted an elitism strategy to speed up the convergence performance, used the non-dominated solutions to add pheromone so as to reinforce the exploitation, and applied a crowding comparison operator to maintain the diversity of the solutions. The results showed that the proposed multi-objective approaches achieved similar or better performance than single objective approaches, so it will be interesting to further investigate the use of multi-objective ACO for feature selection in the future.

An interesting finding in ACO approaches is that a large number of the filter works are based on rough set theory [183], [196], [197], [204], [206], where [204] is the only discovered work on ACO for multi-objective filter feature selection. Jensen and Shen [183] first applied ACO to find a small feature subset in rough set to address feature selection problems. Later, Ming [206] proposed a filter algorithm to use the core features from the rough set as the starting point of ACO for feature selection. Jensen [196] proposed a filter feature selection model based on ACO and fuzzy-rough theory for classification of web content and complex systems monitoring. The popularity of rough set theory in ACO for feature selection is likely because the rough set based measures are easy to update when adding or removing features during the travel of ants.

In summary, in ACO for feature selection, the proportion of filter approaches is much higher than that in GAs, GP, and PSO for feature selection. The graph representation in ACO is more flexible than the representation in GAs and PSO, but the order of encoding the features as nodes may influence the performance. Building feature subsets through ants traversing nodes is similar to many traditional ways of gradually adding or removing features to a subset, which makes it easy to adopt existing filter measures in ACO for feature selection. However, the graph representation may not scale well to problems with thousands of features, which might be the reason why current ACO approaches focus mainly on relatively small-scale problems. Further, investigating the parameter settings in ACO and the capabilities of ACO for multi-objective feature selection are still open issues.

E. Other EC Techniques for Feature Selection

Table V shows other EC techniques for feature selection, including DE, memetic algorithms, LCSs, ES, ABC, AISs,

TABLE V
CATEGORISATION OF OTHER APPROACHES

	DE	Memetic	Others
Filter	[209], [213]	[214]	LCSs[215]; ES [208];
Wrapper	[48], [210], [17], [216], [217], [218], [207]	[219]	ES [53]; ABC [54], [55], [220], [221], [222], [223]; AISs [56], [57]; EDA [224]; GSA [152], [225]; TS [139], [226]; SA [38], [79]
Combined	[47]	[49], [50], [100], [177], [227], [228], [229], [230], [231]	

estimated distribution algorithm (EDA), gravitational search algorithm (GSA), Tabu search (TS), and SA[2], where only [207], [208] are multi-objective approaches. There are many more works on DE and memetic algorithms than on other algorithms listed in Table V.

DE was introduced to solve feature selection problems in recent years, mainly since 2008. Most of the works focus on improving the search mechanisms of DE, while the representation scheme has also been investigated. Khushaba et al. [47] combined DE with ACO for feature selection, where DE was used to search for the optimal feature subset based on the solutions obtained by ACO. Experiments showed that the proposed algorithm achieved better performance than other traditional feature selection algorithms on EEG brain-computer-interface tasks. Ghosh et al. [17] applied an adaptive DE algorithm to feature selection, where the parameters in DE were self-adapting depending on the problems. The results showed that the proposed algorithms outperformed a GA [65], ACO [199], DE [209], and the combination of ACO and DE [47] on image problems. Khushaba et al. [47], [210] proposed a new representation with each individual encoded as a vector of floating numbers and the dimensionality was the desired number of features. The results showed that the proposed DE algorithm achieved better performance than PSO and a GA on EEG brain-computer-interface tasks. DE has also been applied to multi-objective feature selection [207], which showed that the proposed multi-objective approach obtained better feature subsets than single objective approaches in terms of the classification performance and the number of features. However, DE has not been applied to filter multi-objective feature selection, which is an opportunity for future work. Further, DE has achieved success in large-scale optimisation [211], [212], but it has not been investigated for feature selection with a large number of features, e.g. more than 500 or thousands of features.

Memetic algorithms, which combine population based search (an EC technique) with local search, provide a great opportunity to combine wrapper and filter methods. Therefore, in most memetic based feature selection approaches, an EC technique was used for wrapper feature selection and a local search algorithm was used for filter feature selection. Zhu et al. [49], [227], [228] proposed memetic algorithms for feature selection, i.e. GAs for wrapper feature selection and a local search using Markov blanket for filter feature selection. Similarly, local search for filter feature selection using mutual information was applied together with GAs and PSO for wrapper feature selection to develop memetic approaches in [50], [177] and [229]. A two-stage feature selection algorithm was proposed in [214], where a Relief-F algorithm was used to rank individual features and then the top-ranked features were used as input to the memetic wrapper feature selection algorithm. In addition, a memetic algorithm was used for feature selection to improve the performance of LCSs in [231], where a LCS was used as a classification algorithm to evaluate the fitness of the selected features.

Other EC techniques have also been applied to feature selection, mainly including LCSs, ES, ABC, AISs, GSAs, EDAs, TS, and SA. Some of them were combined with other EC techniques [38], [53], [139] while most were applied individually to address feature selection problems [232], [54], [56], [79], [152], [220], [221], [222], [224], [225], [226]. Almost all of them are wrapper based methods.

In summary, a variety of EC techniques have recently been applied to address feature selection problems. Since all algorithm have their own advantages and disadvantages, they can be used for potential further investigation to address different new challenges in the feature selection area.

IV. MEASURES IN FILTER APPROACHES

Feature selection measures have previously been classified into five categories [1]: information measures, consistency measures, dependency (or correlation) measures, distance measures, and precision measures (i.e. wrapper approaches). As this section aims to study typical filter measures used in EC for feature selection, only the first four types of filter measures are reviewed. Since rough set theory and fuzzy set theory are important feature selection measures in computational intelligence, they are also listed as another two separate categories. The six categories of filter measures in EC for feature selection can be seen from Table VI.

Information theory based measures are used more often than all other measures. The use of information measures is mainly in four ways: (1) Use an information measure to rank individual features before using an EC technique. Symmetrical uncertainty or mutual information was used for filter feature ranking and then the top-ranked features were used in ACO [202] or a GA [72] based wrapper feature selection. (2) Use an information measure in the local search of a memetic algorithm. Mutual information [177], [227], symmetrical uncertainty [50], and Markov blanket [228] were used in local search to perform a filter feature selection to refine the solutions obtained by a GA or PSO for wrapper feature selection. (3) Incorporate an information measure into the updating/search mechanism. Mutual information was incorporated in the position updating procedure of PSO in [33] to help improve the performance of PSO and an SVM for wrapper feature selection. Based on information theory and GP, a new relevance measure was proposed in [41] to improve the feature selection and classification performance of GP. (4) Use information theory to form a fitness function in an EC algorithm. This is considered the most popular way to use information theory for feature selection. Based on the idea of "Max-relevance and min-redundancy" [32],

[2]TS and SA are not EC techniques, but we include them here since they have often been used together or compared with EC algorithms.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TEVC.2015.2504420

TABLE VI
MEASURES IN FILTER APPROACHES

Measure	References
Information Theory	[33], [34], [41], [50], [55], [72], [102], [105], [106], [116], [118] [173], [177], [195], [201], [202], [203], [205], [229]
Correlation Measure	[75], [107], [116], [203], [208]
Distance Measure	[11], [15], [118], [136], [171]
Consistency Measure	[98], [105], [166], [203]
Fuzzy Set Theory	[99], [104], [164], [166], [169]
Rough Set Theory	[103], [163], [167], [168], [170], [172], [174], [183], [197], [198], [200], [204]

mutual information was used to measure the redundancy within a feature subset and the relevance between features and the class labels. Different EC methods have been used to maximise the relevance and minimise the redundancy in both single objective and multi-objective manners [34], [102], [106], [195], [201]. However, most of these measures evaluate features individually except for [41], [34], [102].

Correlation measures qualify the ability to predict the value of one variable based on the value of another. Two correlation measures were proposed in [208] to evaluate the relevance and redundancy in ES and NSGAII for feature selection on two credit approval datasets. Li et al. [75] proposed a multiple populations based GA for feature selection, and the correlation between features and the class labels were used as a filter measure to test the performance of the proposed GA.

Distance measures are also known as separability, divergence, or discrimination measures. Iswandy and Koenig [136] used two distance measures, the overlap measure and the compact measure, in PSO for feature selection and successfully reduced the dimensionality. Signal-to-noise ratio was also used for feature selection in [15], where GP was used for classification and the features used by GP were ranked by signal-to-noise ratio with only the top-ranked ones being selected. Signal-to-noise ratio was also used to evaluate the goodness of each individual feature in PSO for feature selection [171].

Consistency measures are based on whether two instances, which have the same feature values, have the same class label. GAs were the first EC technique to use consistency measures [98]. Later, a fuzzy set based consistency measure was proposed in [166], which was different from most consistency measures that required discrete data. The proposed measure was used in PSO for feature selection and shown to be faster than PSO with a fuzzy set based fitness function [164]. Consistency measures are in general computationally more expensive than other filter measures [1], which presents an opportunity for further improvement.

Fuzzy set theory is able to measure imprecision and uncertainty through a membership function, which can be used to evaluate the quality of features. Both PSO and GAs have been used together with a fuzzy fitness function for feature selection in both single objective [99], [164] and multi-objective approaches [104]. Fuzzy set theory has been extensively used for feature selection in non-EC methods and there is still great potential to utilise it in EC based approaches.

Rough set theory can deal with uncertainty and incompleteness. It measures the consistency degree of a dataset through the concept of approximations of a target set, which can be used for feature selection [233]. Wang et al. [163] applied standard rough set theory to form a fitness function in PSO for feature selection. Later, Cervante et al. [167], [170] further used probabilistic rough set theory in PSO for feature selection and achieved better performance than using standard rough set theory. Rough set theory has attracted much attention in ACO for feature selection [183], [196], [204], [206], which has been discussed in Section III-D. The use of rough set was further extended for multi-objective feature selection in GAs [103], PSO [172], [174], and ACO [204] to obtain a set of trade-off feature subsets to better solve the problems. However, most of the existing approaches focus mainly on datasets with a relatively small number of features, say, less than 100.

Using multiple measures simultaneously in a single feature selection algorithm has become popular in recent years since each measure has its own advantages and disadvantages. Spolaôr et al. [105] investigated five different filter measures in NSGAII for feature selection, including inconsistent example pairs as a consistency measure, attribute-class correlation as a dependency/correlation measure, inter-class distance measure, Laplacian score distance measure, and representation entropy as an information measure. The results showed that the combination of the inter-class distance measure and the attribute-class correlation measure performed better than other combinations. Sandin et al. [116] proposed to use information gain, χ^2, odds-ratio, and correlation coefficient, and Ahmed et al. [118] proposed to use information gain and Relief-F to rank individual features. Only the top-ranked features from each measure were used as the input to GP for feature selection. Tallón-Ballesteros and Riquelme [203] tested a correlation measure, a consistency measure, and their combination with information gain in ACO for feature selection. These works show that using multiple measures can help discover useful information in features and improve the performance.

In summary, different types of filter measures have been adopted in EC for feature selection. Among these measures, information measures, correlation measures, and distance measures are computationally relatively cheap while consistency, rough set, and fuzzy set theories based measures may handle noisy data better. However, almost all of them were designed for discrete data and the performance may deteriorate when applied to continuous data, which appears in many real-word problems. It is worth noting that almost all these measures are existing ones (or with little modification), i.e. they were originally used in traditional feature selection methods, e.g. sequential search. EC techniques were used as a search method in these approaches. There are also some measures that are not suitable for using in EC for feature selection because they are designed for a specific (traditional) search method. There are only a few filter measures particularly designed for EC based feature selection, where an example is from Neshatian and Zhang [130] who developed a filter relevance measure based on GP trees with a virtual structure, which improved the performance of GP for feature selection. Compared with wrapper approaches, the classification performance of filter

TABLE VII
APPLICATIONS

Category	Applications	References
(1)	Image analysis	[17], [65], [66], [77], [80], [199], [217]
(1)	Face recognition	[43], [68], [95], [108], [165], [181], [187], [235]
(1)	Music instrument recognition	[53], [93]
(1)	Handwritten digit recognition	[90]
(1)	EEG brain-computer-interface	[47], [210]
(1)	Speaker recognition	[62], [67]
(1)	Personal identification	[137], [213]
(1)	Human action recognition	[84], [201]
(2)	Gene analysis	[49], [103], [15], [118], [119], [122], [139], [140], [143], [145], [171], [189]
(2)	Disease diagnosis	[48], [80], [147]
(3)	Financial problems	[54], [73], [87], [134], [208]
(3)	Customer churn prediction	[92]
(4)	Text mining	[16]
(4)	Web service	[141]
(4)	Network security	[97], [184], [220]
(4)	Email Spam detection	[109]
(5)	Power system optimisation	[79], [152]
(5)	Weed recognition in agriculture	[191]
(5)	Melting point prediction in chemistry	[188]
(5)	Weather forecast	[86], [195]

approaches is usually worse, but they can be much cheaper than wrapper approaches [234], which is critical in large datasets. Therefore, developing filter measures specifically according to the characteristics of an EC technique may significantly increase the efficiency and effectiveness, which offers an important future research direction.

V. APPLICATIONS

Table VII shows the applications of EC for feature selection. It can be seen that EC based feature selection approaches have been applied to a variety of areas.

Generally, the major applications can be grouped into the following five categories: (1) Image and signal processing including image analysis, face recognition, human action recognition, EEG brain-computer-interface, speaker recognition, handwritten digit recognition, personal identification, and music instrument recognition. (2) Biological and biomedical tasks including gene analysis, biomarker detection, and disease diagnosis, where selecting the key features and reducing the dimensionality can significantly reduce the cost of clinic validation, disease diagnosis and other related procedures. (3) Business and financial problems including financial crisis, credit card issuing in bank systems, and customer churn prediction. (4) Network/web service including text mining, web service, network security, and email spam detection. (5) Others, such as power system optimisation, weed recognition in agriculture, melting point prediction in chemistry, and weather prediction.

All the above areas are important and essential to our society or daily life. Of course, many other fields [236], such as complex engineering tasks and language learning, also need feature selection, but EC based approaches have not been thoroughly investigated in those areas.

VI. ISSUES AND CHALLENGES

Despiting the suitability, success and promise of EC for feature selection, there are still significant issues and challenges, which will be discussed here.

A. Scalability

The most pressing issue is due to the trend in "big data" [13], the size of the data becomes increasingly large. In 1989, selecting features from a dataset with more than 20 features was called large-scale feature selection [37]. However, nowadays the number of features in many areas, such as gene analysis, can easily reach thousands or even millions. This increases computational cost and requires advanced search mechanisms, but both of these aspects also have their own issues so the problem cannot be solved by only increasing computational power. Novel methods and algorithms will become necessity.

A number of EC algorithms have been proposed to solve large-scale feature selection problems [70], [72], [15], [118], [140], [176], [202], where the dimensionality reaches a few thousands or tens of thousands. Other computational intelligence based techniques have been introduced to feature selection tasks in the ranges of millions [13], [36]. Most of the existing EC based large-scale feature selection approaches employ a two-stage approach, where in the first stage, a measure is used to evaluate the relevance of individual features, then ranks them according to the relevance value. Only the top-ranked (better) features are used as inputs to the second stage to further select features from them. However, the first stage removes lowly-ranked features without considering their interaction with other features. To solve large-scale feature selection problems, new approaches are needed, including new search algorithms and new evaluation measures. EC approaches have shown their potential for large-scale (global) optimisation [211], [212], [237], which provides a good opportunity to better address large-scale feature selection tasks.

B. Computational Cost

Most feature selection methods suffer from the problem of being computationally expensive, which is a particularly serious issue in EC for feature selection since they often involve a large number of evaluations. Filter approaches are generally more efficient than wrapper approaches, but experiments have shown that this is not always true [234]. Some filter measures, such as the rough set theory [28], [163], [168], [183], [196], [197], [204], [206], may take a longer time than a fast/simple wrapper method [234]. Although there exist fast filter measures, such as mutual information [32], [33], [34], [238], the classification performance is usually worse than most wrapper approaches. Therefore, it is still a challenge to propose efficient and effective approaches to feature selection problems.

To reduce the computational cost, two main factors, an efficient search technique and a fast evaluation measure, need to be considered [1]. A fast evaluation criterion may produce a greater influence than the search technique, since in current approaches the evaluation procedure takes the majority of the

computational cost. It is noted that the parallelisable nature of EC is suited as Grid computing, GPU, and Cloud computing that can be used to speed up the process.

C. Search Mechanisms

Feature selection is an NP-hard problem and has a large complex solution space [239]. This requires a powerful global search technique and current EC algorithms still have great potential to be improved.

The new search mechanisms should have the ability to explore the whole search space and also be able to exploit the local regions when needed. New search mechanisms may involve local search (to form novel memetic algorithms), hybridisation of different EC search mechanisms, hybridisation of EC and conventional methods [39], [158], surrogate approaches [240], etc.

A related issue is that the new search mechanisms should be stable on feature selection tasks. EC algorithms are stochastic approaches, which may produce different solutions when using different starting points. Even when the fitness values of the solutions are the same, they may select different individual features. Therefore, the stability of the algorithms not only involves the difference of the fitness values, but also involves the consistency of the selected features. Therefore, to propose new search algorithms with high stability is also an important task.

D. Measures

The evaluation measure, which forms the fitness function, is one of the key factors in EC for feature selection. It considerably influences the computational time, the classification performance, and the landscape of the search space.

Most of the computational time is spent on the evaluation procedure for wrapper approaches and also for many filter approaches [29], [158], [234]. Although there are some existing fast evaluation measures, such as mutual information [32], [34], [241], [12], they evaluate features individually rather than a group of features. Ignoring interactions between features results in subsets with redundancy and lack of complimentary features [2], [242], which in turn cannot achieve optimal classification performance in most domains of interest. However, discovering complex feature interaction is very challenging and only a few works have been conducted on this direction [243]. There are some measures that can evaluate groups of features [27], [31], [163], [174], but they are usually computationally expensive, such as rough set based measures [163], [174]. Furthermore, many studies show that filter methods do not scale well above tens of thousands of features [13]. Therefore, new measures still need to be developed for feature selection, especially when dealing with large-scale problems.

For feature selection problems, multiple different solutions may have the same fitness values. A small (big) change in the solution may cause a huge (small) difference in the fitness value. This makes the problem even more challenging. Therefore, developing new measures that can smooth the fitness landscape will significantly reduce the difficulty of the task and help the design of suitable search algorithms.

E. Representation

The traditional representation in most EC approaches results in a huge search space for feature selection problems, i.e. the size is 2^n for a dataset with n features, even when n is only a few hundreds [1], [2].

A good representation scheme can help to reduce the search space size. It in turn helps to design new search mechanisms to improve the search ability. Another issue is that the current representations usually reflect only whether a feature is selected or not, but the feature interaction information is not shown. Feature interaction usually involves a group of features rather than a single feature. If the representation can reflect the selection or removal of groups of features, it may significantly improve the classification performance. Furthermore, the interpretation of the solution is also an important issue closely related to the representation. Most EC methods are not good at this task except for GP and LCSs as they produce a tree or a population of rules, which are easier to understand and interpret. Therefore, a good representation scheme may help users better understand and interpret the obtained solutions.

F. Multi-Objective Feature Selection

Most of the existing evolutionary multi-objective (EMO) algorithms are designed for continuous problems [244], but feature selection is a discrete problem. When dealing with large-scale problems, existing EMO methods do not scale well [211], [212], [245], [246]. This requires the development of novel EMO algorithms. Furthermore, the two main objectives (minimising both the number of features and the classification error rate) are not always conflicting with each other, i.e. in some subspaces, decreasing the number of features can also decrease the classification error rate as unnecessary features are removed [29], [154], [158], [171], [173], [194]. This makes it tricky to design an appropriate EMO algorithm. Furthermore, developing new evaluation metrics and further selection methods to choose a single solution from a set of trade-off solutions is also a challenging topic.

Finally, besides the two main objectives, other objectives, such as the complexity, the computational time, and the solution size (e.g. tree size in GP and number of rules in LCSs), could also be considered in multi-objective feature selection.

G. Feature Construction

Feature selection does not create new features, as it only selects original features. However, if the original features are not informative enough to achieve promising performance, feature selection may not work well, yet feature construction may work well [3], [247].

One of the challenges for feature construction is to decide when feature construction is needed. A measure to estimate the properties of the data might be needed to make such a decision. Meanwhile, feature selection and feature construction can be used together to improve the classification performance and reduce the dimensionality. This can be achieved in three different ways: performing feature selection before feature

construction, performing feature construction before feature selection, and simultaneously performing both feature selection and construction [3].

H. Number of Instances

The number of instances in a dataset significantly influences the performance and design of experiments [236]. It causes problems when the number is too big or too small.

When the number of instances is too small, it is hard to design appropriate experiments to test the performance of the algorithms. For example, there might be tens of thousands of features, but the number of instances can be smaller than one hundred because of the high cost of collecting such instances [117]. It is difficult to split the data into a training set and a test set to represent the actual problem. Therefore, many existing works have the problem of feature selection bias [248], [249], especially when the whole set of data is used during the feature selection process [44], [70], [117], [145], [189], [229], [215]. Although cross-validation or bootstrap sampling techniques [250] can address the issue to some extent, they may have the problem of it being hard to decide the final selection of individual features because EC algorithms (and conventional deterministic algorithms) often select different features from different cross-validation runs.

When the number of instances is too big, one major problem is the computational cost [29], [236]. In feature selection, each evaluation usually needs to visit all the training examples. The larger the data/training size, the longer each evaluation. Meanwhile, for "big data" problems, it not only needs to reduce the number of features, but also needs to reduce the number of instances [251]. Combining feature selection and instance selection into a single process may improve both the effectiveness and efficiency of the data pre-processing process.

VII. Conclusions

This paper provided a comprehensive survey of EC techniques in solving feature selection problems, which covered all the commonly used EC algorithms and focused on the key factors, such as representation, search mechanisms, and the performance measures as well as the applications. Important issues and challenges were also discussed.

This survey shows that a variety of EC algorithms have recently attracted much attention to address feature selection tasks. A popular approach in GAs, GP and PSO is to improve the representation to simultaneously select features and optimise the classifiers, e.g. SVMs. Different algorithms have their own characteristics, such as GAs are able to preserve a small set of features during the evolutionary process because of the nature of genetic operators, PSO is relatively computationally cheap because of its simple updating mechanisms, ACO can gradually add features because of the graph representation, and GP can implicitly perform feature selection through feature construction. Therefore, these EC techniques or their combinations can be used with different measures to solve different types of feature selection problems. This needs further investigation in the future. Furthermore, all the major EC algorithms, e.g. GAs, GP and PSO, have been used to address feature selection tasks with thousands of features, but they suffer from the problem of high computational cost. As a result, when are applied to large-scale feature selection tasks, the current target datasets usually have a small number of instances.

Although EC techniques for feature selection have achieved some success, they still face challenges and their potential has not been fully investigated. Scalability is one of the most important issues since both the number of features and the number of instances are increasing in many real-world tasks. This is not only a challenging task in EC, but also in the machine learning, statistics, and biology communities. The recent advances in EC for large-scale global optimisation motivate further studies on EC for large-scale feature selection, but it is challenging to develop promising approaches, where novel search mechanisms and representation schemes are needed in both single objective and multi-objective feature selection. To improve their effectiveness and efficiency, it is necessary to design a cheap evaluation measure according to the specific representation and the search mechanism of a particular EC technique. The proposal of novel approaches may involve methods or measures from different areas, which encourages research across multiple disciplines. A comprehensive comparison between EC and non-EC approaches on a large number of benchmark datasets/problems to test their advantages and disadvantages can help develop novel effective approaches to different kinds of problems. In addition, combining feature selection with feature construction can potentially improve the classification performance while combining feature selection with instance selection can potentially improve the efficiency.

Acknowledgment

This work was supported in part by the Marsden Fund of the New Zealand Government under contract VUW1209, administrated by the Royal Society of New Zealand, and the University Research Fund (210375/3557, 209861/3580) at Victoria University of Wellington. This work was partially supported by EPSRC (Grant No. EP/J017515/1) and NSFC (Grant No. 61329302). Xin Yao was also supported by a Royal Society Wolfson Research Merit Award.

References

[1] M. Dash and H. Liu, "Feature selection for classification," *Intelligent Data Analysis*, vol. 1, no. 4, pp. 131–156, 1997.
[2] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.
[3] A. S. Uday Kamath, Kenneth De Jong, "Effective automated feature construction and selection for classification of biological sequences," *PlosOne*, vol. 9, pp. e99982:1–14, 2014.
[4] W. Albukhanajer, Y. Jin, J. Briffa, and G. Williams, "Evolutionary multi-objective optimization of trace transform for invariant feature extraction," in *IEEE Congress on Evolutionary Computation (CEC)*, pp. 1–8, 2012.
[5] W. Albukhanajer, J. Briffa, and Y. Jin, "Evolutionary multiobjective image feature extraction in the presence of noise," *IEEE Transactions on Cybernetics*, vol. 45, no. 9, pp. 1757–1768, 2015.
[6] Y. Liu, F. Tang, and Z. Zeng, "Feature selection based on dependency margin," *IEEE Transactions on Cybernetics*, vol. 45, no. 6, pp. 1209–1221, 2015.
[7] H. Liu and Z. Zhao, "Manipulating data and dimension reduction methods: Feature selection," in *Encyclopedia of Complexity and Systems Science*, pp. 5348–5359, Springer, 2009.

[8] H. Liu, H. Motoda, R. Setiono, and Z. Zhao, "Feature selection: An ever evolving frontier in data mining," in *Feature Selection for Data Mining*, vol. 10 of *JMLR Proceedings*, pp. 4–13, JMLR.org, 2010.

[9] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 4, pp. 491–502, 2005.

[10] A. Unler and A. Murat, "A discrete particle swarm optimization method for feature selection in binary classification problems," *European Journal of Operational Research*, vol. 206, no. 3, pp. 528–539, 2010.

[11] Y. Liu, G. Wang, H. Chen, and H. Dong, "An improved particle swarm optimization for feature selection," *Journal of Bionic Engineering*, vol. 8, no. 2, pp. 191–200, 2011.

[12] J. Vergara and P. Estévez, "A review of feature selection methods based on mutual information," *Neural Computing and Applications*, vol. 24, no. 1, pp. 175–186, 2014.

[13] Y. Zhai, Y.-S. Ong, and I. Tsang, "The emerging "big dimensionality"," *IEEE Computational Intelligence Magazine*, vol. 9, no. 3, pp. 14–26, 2014.

[14] B. de la Iglesia, "Evolutionary computation for feature selection in classification problems," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 3, no. 6, pp. 381–407, 2013.

[15] S. Ahmed, M. Zhang, and L. Peng, "Enhanced feature selection for biomarker discovery in LC-MS data using GP," in *IEEE Congress on Evolutionary Computation (CEC)*, pp. 584–591, 2013.

[16] M. H. Aghdam, N. Ghasem-Aghaee, and M. E. Basiri, "Text feature selection using ant colony optimization," *Expert Systems with Applications*, vol. 36, no. 3, Part 2, pp. 6843–6853, 2009.

[17] A. Ghosh, A. Datta, and S. Ghosh, "Self-adaptive differential evolution for feature selection in hyperspectral image data," *Applied Soft Computing*, vol. 3, pp. 1969–1977, 2013.

[18] K. Krawiec, "Evolutionary feature selection and construction," in *Encyclopedia of Machine Learning*, pp. 353–357, Springer US, 2010.

[19] D. Muni, N. Pal, and J. Das, "Genetic programming for simultaneous feature selection and classifier design," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 36, no. 1, pp. 106–117, 2006.

[20] L. Jung-Yi, K. Hao-Ren, C. Been-Chian, and Y. Wei-Pang, "Classifier design with feature selection and feature extraction using layered genetic programming," *Expert Systems with Applications*, vol. 34, no. 2, pp. 1384–1393, 2008.

[21] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, pp. 273–324, 1997.

[22] M. Mitchell, *An Introduction to Genetic Algorithms*. The MIT Press, 1996.

[23] A. Whitney, "A direct method of nonparametric measurement selection," *IEEE Transactions on Computers*, vol. C-20, no. 9, pp. 1100–1103, 1971.

[24] T. Marill and D. Green, "On the effectiveness of receptors in recognition systems," *IEEE Transactions on Information Theory*, vol. 9, no. 1, pp. 11–17, 1963.

[25] S. Stearns, "On selecting features for pattern classifier," in *Proceedings of the 3rd International Conference on Pattern Recognition*, (Coronado, Calif, USA), pp. 71–75, IEEE Press, 1976.

[26] P. Pudil, J. Novovicova, and J. V. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, vol. 15, no. 11, pp. 1119–1125, 1994.

[27] Q. Mao and I.-H. Tsang, "A feature selection method for multivariate performance measures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 9, pp. 2051–2063, 2013.

[28] F. Min, Q. Hu, and W. Zhu, "Feature selection with test cost constraint," *International Journal of Approximate Reasoning*, vol. 55, no. 1, Part 2, pp. 167–179, 2014.

[29] B. Xue, M. Zhang, and W. N. Browne, "Particle swarm optimization for feature selection in classification: A multi-objective approach," *IEEE Transactions on Cybernetics*, vol. 43, no. 6, pp. 1656–1671, 2013.

[30] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Proceedings of the ninth international workshop on Machine learning*, pp. 249–256, 1992.

[31] S. Wang, W. Pedrycz, Q. Zhu, and W. Zhu, "Subspace learning for unsupervised feature selection via matrix factorization," *Pattern Recognition*, vol. 48, no. 1, pp. 10–19, 2015.

[32] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[33] A. Unler and R. B. C. Alper Murat, "mr2PSO: A maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification," *Information Sciences*, vol. 20, pp. 4625–4641, 2011.

[34] L. Cervante, B. Xue, M. Zhang, and L. Shang, "Binary particle swarm optimisation for feature selection: A filter based approach," in *IEEE Congress on Evolutionary Computation (CEC)*, pp. 881–888, 2012.

[35] N. C. Tan, W. G. Fisher, K. P. Rosenblatt, and H. R. Garner, "Application of multiple statistical tests to enhance mass spectrometry-based biomarker discovery," *BMC bioinformatics*, vol. 10, no. 1, p. 144, 2009.

[36] M. Tan, I. Tsang, and L. Wang, "Minimax sparse logistic regression for very high-dimensional feature selection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 10, pp. 1609–1622, 2013.

[37] W. Siedlecki and J. Sklansky, "A note on genetic algorithms for large-scale feature selection," *Pattern Recognition Letters*, vol. 10, no. 5, pp. 335–347, 1989.

[38] I. A. Gheyas and L. S. Smith, "Feature subset selection in large dimensionality domains," *Pattern Recognition*, vol. 43, no. 1, pp. 5–13, 2010.

[39] Y. Jeong, K. S. Shin, and M. K. Jeong, "An evolutionary algorithm with the partial sequential forward floating search mutation for large-scale feature selection problems," *Journal of the Operational Research Society*, pp. 1–10, 2014.

[40] A. Purohit, N. Chaudhari, and A. Tiwari, "Construction of classifier with feature selection based on genetic programming," in *IEEE Congress on Evolutionary Computation (CEC)*, pp. 1–5, 2010.

[41] K. Neshatian and M. Zhang, "Improving relevance measures using genetic programming," in *European Conference on Genetic Programming (EuroGP 2012)*, vol. 7244 of *Lecture Notes in Computer Science*, pp. 97–108, Springer, 2012.

[42] D. K. Agrafiotis and W. Cedeño, "Feature selection for structure-activity correlation using binary particle swarms," *Journal of Medicinal Chemistry*, vol. 45, no. 5, pp. 1098–1107, 2002.

[43] Z. Yan and C. Yuan, "Ant colony optimization for feature selection in face recognition," in *Biometric Authentication*, vol. 3072 of *Lecture Notes in Computer Science*, pp. 221–226, Heidelberg, 2004.

[44] L. Santana, L. Silva, A. Canuto, F. Pintro, and K. Vale, "A comparative analysis of genetic algorithm and ant colony optimization to select attributes for an heterogeneous ensemble of classifiers," in *IEEE Congress on Evolutionary Computation*, pp. 1–8, 2010.

[45] B. Chen, L. Chen, and Y. Chen, "Efficient ant colony optimization for image feature selection," *Signal Processing*, vol. 93, pp. 1566–1576, 2012.

[46] X. Zhao, D. Li, B. Yang, C. Ma, Y. Zhu, and H. Chen, "Feature selection based on improved ant colony optimization for online detection of foreign fiber in cotton," *Applied Soft Computing*, vol. 24, pp. 585–596, 2014.

[47] R. Khushaba, A. Al-Ani, A. AlSukker, and A. Al-Jumaily, "A combined ant colony and differential evolution feature selection algorithm," in *Ant Colony Optimization and Swarm Intelligence*, vol. 5217 of *Lecture Notes in Computer Science*, pp. 1–12, Heidelberg, 2008.

[48] L. Wang, H. Ni, R. Yang, V. Pappu, M. B. Fenn, and P. M. Pardalos *Optimization Methods and Software*, vol. 29, no. 4, pp. 703–719, 2014.

[49] Z. Zhu, Y.-S. Ong, and M. Dash, "Markov blanket-embedded genetic algorithm for gene selection," *Pattern Recognition*, vol. 40, no. 11, pp. 3236–3248, 2007.

[50] S. S. Kannan and N. Ramaraj, "A novel hybrid feature selection via symmetrical uncertainty ranking based local memetic search algorithm," *Knowledge-Based Systems*, vol. 23, no. 6, pp. 580–585, 2010.

[51] M. Iqbal, S. S. Naqvi, W. N. Browne, C. Hollitt, and M. Zhang, "Salient object detection using learning classifier systems that compute action mappings," in *Proceedings of the 16th Annual Conference on Genetic and Evolutionary Computation (GECCO)*, pp. 525–532, 2014.

[52] I. M. Alvarez, W. N. Browne, and M. Zhang, "Reusing learned functionality in xcs: Code fragments with constructed functionality and constructed features," in *the 2014 Conference Companion on Genetic and Evolutionary Computation Companion (GECCO)*, pp. 969–976, 2014.

[53] I. Vatolkin, W. Theimer, and G. Rudolph, "Design and comparison of different evolution strategies for feature selection and consolidation in music classification," in *IEEE Congress on Evolutionary Computation*, pp. 174–181, 2009.

[54] M. Marinaki and Y. Marinakis, "A bumble bees mating optimization algorithm for the feature selection problem," *International Journal of Machine Learning and Cybernetics*, pp. 1–20 (Published online), 2014.

[55] E. Hancer, B. Xue, M. Zhang, D. Karaboga, and B. Akay, "A multi-objective artificial bee colony approach to feature selection using fuzzy

mutual information," in *Evolutionary Computation (CEC), 2015 IEEE Congress on*, pp. 2420–2427, May 2015.

[56] S.-W. Lin and S.-C. Chen, "Parameter tuning, feature selection and weight assignment of features for case-based reasoning by artificial immune system," *Applied Soft Computing*, vol. 11, no. 8, pp. 5042–5052, 2011.

[57] K.-J. Wang, K.-H. Chen, and M.-A. Angelia, "An improved artificial immune recognition system with the opposite sign test for feature selection," *Knowledge-Based Systems*, vol. 71, pp. 126–145, 2014.

[58] M. Komosiski and K. Krawiec, "Evolutionary weighting of image features for diagnosing of CNS tumors," *Artificial Intelligence in Medicine*, vol. 19, no. 1, pp. 25–38, 2000.

[59] R. Leardi, R. Boggia, and M. Terrile, "Genetic algorithms as a strategy for feature selection," *Journal of Chemometrics*, vol. 6, no. 5, pp. 267–281, 1992.

[60] H. Vafaie and K. DeJong, "Feature space transformation using genetic algorithms," *IEEE Intelligent Systems*, vol. 13, no. 2, pp. 57–65, 1998.

[61] J. Yang and V. Honavar, "Feature subset selection using a genetic algorithm," *IEEE Intelligent Systems and their Applications*, vol. 13, no. 2, pp. 44–49, 1998.

[62] M. Demirekler and A. Haydar, "Feature selection using genetics-based algorithm and its application to speaker identification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 329–332, 1999.

[63] T. N. Hisao Ishibuchi, "Multi-objective pattern and feature selection by a genetic algorithm," in *Proceeding of the 2nd Annual Conference on Genetic and Evolutionary Computation Conference (GECCO)*, pp. 1069–1076, 2000.

[64] M. G. Smith and L. Bull, "Genetic programming with a genetic algorithm for feature construction and selection," *Genetic Programming and Evolvable Machines*, vol. 6, no. 3, pp. 265–281, 2005.

[65] P. S. Shixin Yu, Steve De Backer, "Genetic feature selection combined with composite fuzzy nearest neighbor classifiers for hyperspectral satellite imagery," *Pattern Recognition Letters*, vol. 23, no. 1–3, pp. 183–190, 2002.

[66] K. Umamaheswari, S. Sumathi, and S. N. Sivanandam, "Neuro - genetic approaches to classification of face images with effective feature selection using hybrid classifiers," in *International Conference on Advanced Computing and Communications*, pp. 286–291, 2006.

[67] M. Zamalloa, G. Bordel, L. Rodriguez, and M. Penagarikano, "Feature selection based on genetic algorithms for speaker recognition," in *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The*, pp. 1–8, 2006.

[68] M. Shoorehdeli, M. Teshnehlab, and H. Moghaddam, "Feature subset selection for face detection using genetic algorithms and particle swarm optimization," in *IEEE International Conference on Networking, Sensing and Control (ICNSC)*, pp. 686–690, 2006.

[69] J.-H. Hong and S.-B. Cho, "Efficient huge-scale feature selection with speciated genetic algorithm," *Pattern Recognition Letters*, vol. 27, no. 2, pp. 143 – 150, 2006.

[70] E. Alba, J. Garcia-Nieto, and L. Jourdan, "Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms," in *IEEE Congress on Evolutionary Computation (CEC)*, pp. 284–290, 2007.

[71] D. Chen, K. Chan, and X. Wu, "Gene expression analyses using genetic algorithm based hybrid approaches," in *IEEE Congress on Evolutionary Computation*, pp. 963–969, 2008.

[72] F. Tan, X. Z. Fu, Y. Q. Zhang, and A. Bourgeois, "A genetic algorithm-based method for feature subset selection," *Soft Computing*, vol. 12, no. 5, pp. 111–120, 2008.

[73] L.-H. Chen and H.-D. Hsiao, "Feature selection to diagnose a business crisis by using a real ga-based support vector machine: An empirical study," *Expert Systems with Applications*, vol. 35, no. 3, pp. 1145–1155, 2008.

[74] H.-W. Cho, S. B. Kim, M. K. Jeong, Y. Park, T. R. Ziegler, and D. P. Jones, "Genetic algorithm-based feature selection in high-resolution NMR spectra," *Expert Systems with Applications*, vol. 35, no. 3, pp. 967–975, 2008.

[75] Y. Li, S. Zhang, and X. Zeng, "Research of multi-population agent genetic algorithm for feature selection," *Expert Systems with Applications*, vol. 36, no. 9, pp. 11570–11581, 2009.

[76] J. Derrac, S. Garcia, and F. Herrera, "A first study on the use of coevolutionary algorithms for instance and feature selection," *Hybrid Artificial Intelligence Systems, Lecture Notes in Computer Science*, vol. 5572, pp. 557– 564, 2009.

[77] R. Li, J. Lu, Y. Zhang, and T. Zhao, "Dynamic adaboost learning with feature selection based on parallel genetic algorithm for image annotation," *Knowledge-Based Systems*, vol. 23, no. 3, pp. 195–201, 2010.

[78] S. C. Yusta, "Different metaheuristic strategies to solve the feature selection problem," *Pattern Recognition Letters*, vol. 30, pp. 525–534, 2009.

[79] K. Manimala, K. Selvi, and R. Ahila, "Hybrid soft computing techniques for feature selection and parameter optimization in power quality data mining," *Applied Soft Computing*, vol. 11, no. 8, pp. 5485–5497, 2011.

[80] S. F. da Silva, M. X. Ribeiro, J. do E.S. Batista Neto, C. Traina-Jr., and A. J. Traina, "Improving the ranking quality of medical image retrieval using a genetic feature selection method," *Decision Support Systems*, vol. 51, no. 4, pp. 810–820, 2011.

[81] S. M. Winkler, M. Affenzeller, W. Jacak, and H. Stekel, "Identification of cancer diagnosis estimation models using evolutionary algorithms: A case study for breast cancer, melanoma, and cancer in the respiratory system," in *the 13th Annual Conference Companion on Genetic and Evolutionary Computation (GECCO)*, pp. 503–510, ACM, 2011.

[82] T. C. Chen, Y. C. Hsieh, P. S. You, and Y. C. Lee, "Feature selection and classification by using grid computing based evolutionary approach for the microarray data," in *3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT)*, vol. 9, pp. 85–89, 2010.

[83] F. Souza, T. Matias, and R. Araujo, "Co-evolutionary genetic multilayer perceptron for feature selection and model design," in *16th Conference on Emerging Technologies Factory Automation (ETFA)*, pp. 1–7, 2011.

[84] A. A. Chaaraoui and F. Florez-Revuelta, "Human action recognition optimization based on evolutionary feature subset selection," in *Proceedings of the 15th Annual Conference on Genetic and Evolutionary Computation (GECCO)*, pp. 1229–1236, ACM, 2013.

[85] S. Oreski and G. Oreski, "Genetic algorithm-based heuristic for feature selection in credit risk assessment," *Expert Systems with Applications*, vol. 41, no. 4, Part 2, pp. 2052–2064, 2014.

[86] J.-H. Seo, Y. H. Lee, and Y.-H. Kim, "Feature selection for very short-term heavy rainfall prediction using evolutionary computation," *Advances in Meteorology*, pp. 203545:1–15, 2014.

[87] D. Liang, C.-F. Tsai, and H.-T. Wu, "The effect of feature selection on financial distress prediction," *Knowledge-Based Systems*, vol. 73, pp. 289–297, 2015.

[88] C. Emmanouilidis, A. Hunter, and J. MacIntyre, "A multiobjective evolutionary setting for feature selection and a commonality-based crossover operator," in *IEEE Congress on Evolutionary Computation*, vol. 1, pp. 309–316, 2000.

[89] Y. Kim, W. N. Street, and F. Menczer, "Feature selection in unsupervised learning via evolutionary search," in *the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 365–369, 2000.

[90] L. Oliveira, R. Sabourin, F. Bortolozzi, and C. Suen, "Feature selection using multi-objective genetic algorithms for handwritten digit recognition," in *16th International Conference on Pattern Recognition (ICPR)*, vol. 1, pp. 568– 571, 2002.

[91] K. Waqas, R. Baig, and S. Ali, "Feature subset selection using multi-objective genetic algorithms," in *IEEE 13th International Conference on Multitopic Conference (INMIC)*, pp. 1–6, 2009.

[92] B. Huang, B. Buckley, and T.-M. Kechadi, "Multi-objective feature selection by using NSGA-II for customer churn prediction in telecommunications," *Expert Systems with Applications*, vol. 37, pp. 3638–3646, 2010.

[93] I. Vatolkin, M. Preuß, G. Rudolph, M. Eichhoff, and C. Weihs, "Multi-objective evolutionary feature selection for instrument recognition in polyphonic audio mixtures," *Soft Computing*, vol. 16, no. 12, pp. 2027–2047, 2012.

[94] A. Mukhopadhyay and U. Maulik, "An SVM-wrapped multiobjective evolutionary feature selection approach for identifying cancer-microrna markers," *IEEE Transactions on NanoBioscience*, vol. 12, no. 4, pp. 275–281, 2013.

[95] L. D. Vignolo, D. H. Milone, and J. Scharcanskia, "Feature selection for face recognition based on multi-objective evolutionary wrappers," *Expert Systems with Applications*, vol. 40, pp. 5077–5084, 2013.

[96] C. J. Tan, C. P. Lim, and Y.-N. Cheah, "A multi-objective evolutionary algorithm-based ensemble optimizer for feature selection and classification with neural network models," *Neurocomputing*, vol. 125, pp. 217–228, 2014.

[97] E. de la Hoz, E. de la Hoz, A. Ortiz, J. Ortega, and A. Martínez-Álvarez, "Feature selection by multi-objective optimisation: Application to network anomaly detection by hierarchical self-organising maps," *Knowledge-Based Systems*, vol. 71, pp. 322–338, 2014.

[98] P.-L. Lanzi, "Fast feature selection with genetic algorithms: a filter approach," in *IEEE International Conference on Evolutionary Computation*, pp. 537–540, 1997.

[99] B. Chakraborty, "Genetic algorithm with fuzzy fitness function for feature selection," in *IEEE International Symposium on Industrial Electronics (ISIE)*, vol. 1, pp. 315– 319, 2002.

[100] W. Sheng, X. Liu, and M. Fairhurst, "A niching memetic algorithm for simultaneous clustering and feature selection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 7, pp. 868–879, 2008.

[101] A. Saxena, N. Pal, and M. Vora, "Evolutionary methods for unsupervised feature selection using sammon's stress function," *Fuzzy Information and Engineering*, vol. 2, no. 3, pp. 229–247, 2010.

[102] B. Xue, L. Cervante, L. Shang, W. N. Browne, and M. Zhang, "Multi-objective evolutionary algorithms for filter based feature selection in classification," *International Journal on Artificial Intelligence Tools*, vol. 22, no. 04, pp. 1350024:1–31, 2013.

[103] M. Banerjee, S. Mitra, and H. Banka, "Evolutionary rough feature selection in gene expression data," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 37, no. 4, pp. 622–632, 2007.

[104] P. Kundu and S. Mitra, "Multi-objective evolutionary feature selection," in *Pattern Recognition and Machine Intelligence*, vol. 5909 of *Lecture Notes in Computer Science*, pp. 74–79, Heidelberg, 2009.

[105] N. Spolaôr, A. Lorena, and H. Lee, "Multi-objective genetic algorithm evaluation in feature selection," in *Evolutionary Multi-Criterion Optimization*, vol. 6576 of *Lecture Notes in Computer Science*, pp. 462–476, Heidelberg, 2011.

[106] H. Xia, J. Zhuang, and D. Yu, "Multi-objective unsupervised feature selection algorithm utilizing redundancy measure and negative epsilon-dominance for fault diagnosis," *Neurocomputing*, vol. 146, pp. 113–124, 2014.

[107] A. Canuto and D. Nascimento, "A genetic-based approach to features selection for ensembles using a hybrid and adaptive fitness function," in *International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2012.

[108] D. Mazumdar, S. Mitra, and S. Mitra, "Evolutionary-rough feature selection for face recognition," in *Transactions on Rough Sets XII*, vol. 6190 of *Lecture Notes in Computer Science*, pp. 117–142, Heidelberg, 2010.

[109] P. Sousa, P. Cortez, R. Vaz, M. Rocha, and M. Rio, "Email spam detection: A symbiotic feature selection approach fostered by evolutionary computation," *International Journal of Information Technology & Decision Making*, vol. 12, no. 04, pp. 863–884, 2013.

[110] D. Chyzhyk, A. Savio, and M. Graña, "Evolutionary ELM wrapper feature selection for alzheimer's disease CAD on anatomical brain MRI," *Neurocomputing*, vol. 128, pp. 73–80, 2014.

[111] F. Lin, D. Liang, C.-C. Yeh, and J.-C. Huang, "Novel feature selection methods to financial distress prediction," *Expert Systems with Applications*, vol. 41, no. 5, pp. 2472–2483, 2014.

[112] A. A. Yahya, A. Osman, A. R. Ramli, and A. Balola, "Feature selection for high dimensional data: an evolutionary filter approach," *Journal of Computer Science*, vol. 7, no. 5, pp. 800–820, 2011.

[113] K. Krawiec, "Genetic programming-based construction of features for machine learning and knowledge discovery tasks," *Genetic Programming and Evolvable Machines*, vol. 3, pp. 329–343, 2002.

[114] R. A. Davis, A. J. Charlton, S. Oehlschlager, and J. C. Wilson, "Novel feature selection method for genetic programming using metabolomic 1h NMR data," *Chemometrics and Intelligent Laboratory Systems*, vol. 81, no. 1, pp. 50–59, 2006.

[115] R. Hunt, K. Neshatian, and M. Zhang, "A genetic programming approach to hyper-heuristic feature selection," in *Simulated Evolution and Learning*, vol. 7673 of *Lecture Notes in Computer Science*, pp. 320–330, Heidelberg, 2012.

[116] I. Sandin, G. Andrade, F. Viegas, D. Madeira, L. Rocha, T. Salles, and M. Goncalves, "Aggressive and effective feature selection using genetic programming," in *IEEE Congress on Evolutionary Computation (CEC)*, pp. 1–8, 2012.

[117] S. Ahmed, M. Zhang, and L. Peng, "Genetic programming for biomarker detection in mass spectrometry data," in *AI 2012: Advances in Artificial Intelligence*, vol. 7691 of *Lecture Notes in Computer Science*, pp. 266–278, Heidelberg, 2012.

[118] S. Ahmed, M. Zhang, and L. Peng, "Improving feature ranking for biomarker discovery in proteomics mass spectrometry data using genetic programming," *Connection Science*, vol. 26, no. 3, pp. 215–243, 2014.

[119] S. Ahmed, M. Zhang, and L. Peng, "Feature selection and classification of high dimensional mass spectrometry data: A genetic programming approach," in *EvoBio*, vol. 7833 of *Lecture Notes in Computer Science*, pp. 43–55, Heidelberg, 2013.

[120] K. Neshatian and M. Zhang, "Using genetic programming for context-sensitive feature scoring in classification problems," *Connection Science*, vol. 23, no. 3, pp. 183–207, 2011.

[121] A. Friedlander, K. Neshatian, and M. Zhang, "Meta-learning and feature ranking using genetic programming for classification: Variable terminal weighting," in *IEEE Congress on Evolutionary Computation*, pp. 941–948, 2011.

[122] S. Ahmed, M. Zhang, and L. Peng, "Prediction of detectable peptides in ms data using genetic programming," in *the 2014 Conference Companion on Genetic and Evolutionary Computation Companion (GECCO)*, pp. 37–38, 2014.

[123] J. Sherrah, R. E. Bogner, and A. Bouzerdoum, "Automatic selection of features for classification using genetic programming," in *Australian and New Zealand Conference on Intelligent Information Systems*, pp. 284–287, 1996.

[124] V. Venkatraman, A. R. Dalby, and Z. R. Yang, "Evaluation of mutual information and genetic programming for feature selection in qsar," *Journal of Chemical Information and Computer Sciences*, vol. 44, no. 5, pp. 1686–1692, 2004.

[125] B. C. Chien and J. H. Yang, "Features selection based on rough membership and genetic programming," in *IEEE International Conference on Systems, Man and Cybernetics (SMC)*, vol. 5, pp. 4124–4129, 2006.

[126] D. Harvey and M. Todd, "Automated feature design for numeric sequence classification by genetic programming," *IEEE Transactions on Evolutionary Computation*, vol. 19, no. 4, pp. 474–489, 2015.

[127] K. Neshatian and M. Zhang, "Unsupervised elimination of redundant features using genetic programming," in *22nd Australasian Joint Conference on Artificial Intelligence*, vol. 5866 of *Lecture Notes in Computer Science*, pp. 432–442, Springer, 2009.

[128] K. Neshatian and M. Zhang, "Dimensionality reduction in face detection: A genetic programming approach," in *24th International Conference Image and Vision Computing New Zealand (IVCNZ)*, pp. 391–396, IEEE, 2009.

[129] K. Neshatian and M. Zhang, "Pareto front feature selection: using genetic programming to explore feature space," in *Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation (GECCO)*, pp. 1027–1034, 2009.

[130] K. Neshatian and M. Zhang, "Genetic programming for feature subset ranking in binary classification problems," in *European Conference on Genetic Programming*, (Berlin, Heidelberg), pp. 121–132, Springer-Verlag, 2009.

[131] S. Nguyen, M. Zhang, M. Johnston, and K. C. Tan, "Automatic design of scheduling policies for dynamic multi-objective job shop scheduling via cooperative coevolution genetic programming," *IEEE Transactions on Evolutionary Computation*, vol. 18, no. 2, pp. 193–208, 2014.

[132] H. Al-Sahaf, M. Zhang, and M. Johnston, "Genetic programming for multiclass texture classification using a small number of instances," in *Simulated Evolution and Learning*, vol. 8886 of *Lecture Notes in Computer Science*, pp. 335–346, 2014.

[133] I. Babaoglu, O. Findik, and E. Ulker, "A comparison of feature selection models utilizing binary particle swarm optimization and genetic algorithm in determining coronary artery disease using support vector machine," *Expert Systems with Applications*, vol. 37, no. 4, pp. 3177 – 3183, 2010.

[134] C. Zhang and H. Hu, "Using PSO algorithm to evolve an optimum input subset for a SVM in time series forecasting," in *IEEE International Conference on Systems, Man and Cybernetics (SMC)*, vol. 4, pp. 3793–3796, 2005.

[135] E. K. Tang, P. Suganthan, and X. Yao, "Feature selection for microarray data using least squares SVM and particle swarm optimization," in *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pp. 1–8, 2005.

[136] K. Iswandy and A. Koenig, "Feature-level fusion by multi-objective binary particle swarm based unbiased feature selection for optimized sensor system design," in *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pp. 365 –370, 2006.

[137] G. Azevedo, G. Cavalcanti, and E. Filho, "An approach to feature selection for keystroke dynamics systems based on PSO and feature weighting," in *IEEE Congress on Evolutionary Computation (CEC)*, pp. 3577–3584, 2007.

[138] S. W. Lin, K. C. Ying, S. C. Chen, and Z. J. Lee, "Particle swarm optimization for parameter determination and feature selection of support vector machines," *Expert Systems with Applications*, vol. 35, no. 4, pp. 1817–1824, 2008.

[139] Q. Shen, W.-M. Shi, and W. Kong, "Hybrid particle swarm optimization and tabu search approach for selecting genes for tumor classification using gene expression data," *Computational Biology and Chemistry*, vol. 32, no. 1, pp. 52–59, 2008.

[140] L. Y. Chuang, H. W. Chang, C. J. Tu, and C. H. Yang, "Improved binary PSO for feature selection using gene expression data," *Computational Biology and Chemistry*, vol. 32, no. 29, pp. 29– 38, 2008.

[141] C. L. Huang and J. F. Dun, "A distributed PSO-SVM hybrid system with feature selection and parameter optimization," *Application on Soft Computing*, vol. 8, pp. 1381–1391, 2008.

[142] A. Mohemmed, M. Zhang, and M. Johnston, "Particle swarm optimization based adaboost for face detection," in *IEEE Congress on Evolutionary Computation (CEC)*, pp. 2494–2501, 2009.

[143] L. Y. Chuang, H. W. Chang, C. J. Tu, and C. H. Yang, "Improved binary PSO for feature selection using gene expression data," *Computational Biology and Chemistry*, vol. 32, no. 29, pp. 29– 38, 2008.

[144] S.-W. Lin and S.-C. Chen, "Psolda: A particle swarm optimization approach for enhancing classification accuracy rate of linear discriminant analysis," *Appled Soft Computing*, vol. 9, no. 3, pp. 1008–1015, 2009.

[145] M. Mohamad, S. Omatu, S. Deris, and M. Yoshioka, "A modified binary particle swarm optimization for selecting the small subset of informative genes from gene expression data," *IEEE Transactions on Information Technology in Biomedicine*, vol. 15, no. 6, pp. 813–822, 2011.

[146] L. Y. Chuang, S. W. Tsai, and C. H. Yang, "Improved binary particle swarm optimization using catfish effect for feature selection," *Expert Systems with Applications*, vol. 38, pp. 12699–12707, 2011.

[147] S. Vieira, L. Mendonca, G. Farinha, and J. Sousa, "Metaheuristics for feature selection: Application to sepsis outcome prediction," in *IEEE Congress on Evolutionary Computation (CEC)*, pp. 1 –8, 2012.

[148] B. Xue, M. Zhang, and W. N. Browne, "New fitness functions in binary particle swarm optimisation for feature selection," in *IEEE Congress on Evolutionary Computation (CEC)*, pp. 2145–2152, 2012.

[149] B. Xue, M. Zhang, and W. N. Browne, "Single feature ranking and binary particle swarm optimisation based feature subset ranking for feature selection," in *Australasian Computer Science Conference (ACSC 2012)*, vol. 122 of *CRPIT*, pp. 27–36, 2012.

[150] C. Jin, S.-W. Jin, and L.-N. Qin, "Attribute selection method based on a hybrid BPNN and PSO algorithms," *Applied Soft Computing*, vol. 12, no. 8, pp. 2147–2155, 2012.

[151] B. Xue, M. Zhang, and W. Browne, "Novel initialisation and updating mechanisms in PSO for feature selection in classification," in *Applications of Evolutionary Computation*, vol. 7835 of *Lecture Notes in Computer Science*, pp. 428–438, Heidelberg, 2013.

[152] C. Ramos, A. de Souza, A. Falcao, and J. Papa, "New insights on nontechnical losses characterization through evolutionary-based feature selection," *IEEE Transactions on Power Delivery*, vol. 27, no. 1, pp. 140 –146, 2012.

[153] S. M. Vieira, L. F. Mendonça, G. J. Farinha, and J. M. Sousa, "Modified binary PSO for feature selection using SVM applied to mortality prediction of septic patients," *Applied Soft Computing*, vol. 13, no. 5, pp. 3494–3504, 2013.

[154] M. Lane, B. Xue, I. Liu, and M. Zhang, "Particle swarm optimisation and statistical clustering for feature selection," in *AI 2013: Advances in Artificial Intelligence*, vol. 8272 of *Lecture Notes in Computer Science*, pp. 214–220, Springer, 2013.

[155] H. Nguyen, B. Xue, I. Liu, and M. Zhang, "PSO and statistical clustering for feature selection: A new representation," in *Simulated Evolution and Learning*, vol. 8886 of *Lecture Notes in Computer Science*, pp. 569–581, 2014.

[156] B. Tran, B. Xue, and M. Zhang, "Improved pso for feature selection on high-dimensional datasets," in *Simulated Evolution and Learning*, vol. 8886 of *Lecture Notes in Computer Science*, pp. 503–515, 2014.

[157] M. Lane, B. Xue, I. Liu, and M. Zhang, "Gaussian based particle swarm optimisation and statistical clustering for feature selection," in *Evolutionary Computation in Combinatorial Optimisation*, vol. 8600 of *Lecture Notes in Computer Science*, pp. 133–144, 2014.

[158] B. Xue, M. Zhang, and W. N. Browne, "Particle swarm optimisation for feature selection in classification: Novel initialisation and updating mechanisms," *Applied Soft Computing*, vol. 18, pp. 261–276, 2014.

[159] H. Nguyen, B. Xue, I. Liu, P. Andreae, and M. Zhang, "Gaussian transformation based representation in particle swarm optimisation for feature selection," in *Applications of Evolutionary Computation*, vol. 9028 of *Lecture Notes in Computer Science*, pp. 541–553, 2015.

[160] Z. Yong, G. Dunwei, H. Ying, and Z. Wanqiu, "Feature selection algorithm based on bare bones particle swarm optimization," *Neurocomputing*, vol. 148, pp. 150–157, 2015.

[161] B. Xue, M. Zhang, and W. N. Browne, "Multi-objective particle swarm optimisation (PSO) for feature selection," in *Proceeding of the 14th Annual Conference on Genetic and Evolutionary Computation Conference (GECCO)*, pp. 81–88, ACM, 2012.

[162] Y. Zhang, C. Xia, D. Gong, and X. Sun, "Multi-objective PSO algorithm for feature selection problems with unreliable data," in *Advances in Swarm Intelligence*, vol. 8794 of *Lecture Notes in Computer Science*, pp. 386–393, Springer, 2014.

[163] X. Wang, J. Yang, X. Teng, W. Xia, and R. Jensen, "Feature selection based on rough sets and particle swarm optimization," *Pattern Recognition Letters*, vol. 28, no. 4, pp. 459–471, 2007.

[164] B. Chakraborty, "Feature subset selection by particle swarm optimization with fuzzy fitness function," in *3rd International Conference on Intelligent System and Knowledge Engineering (ISKE)*, vol. 1, pp. 1038–1042, 2008.

[165] M. Aneesh, A. A. Masand, and K. Manikantan, "Optimal feature selection based on image pre-processing using accelerated binary particle swarm optimization for enhanced face recognition," *Procedia Engineering*, vol. 30, no. 5, pp. 750–758, 2012.

[166] B. Chakraborty and G. Chakraborty, "Fuzzy consistency measure with particle swarm optimization for feature selection," in *2013 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 4311–4315, 2013.

[167] L. Cervante, B. Xue, L. Shang, and M. Zhang, "Binary particle swarm optimisation and rough set theory for dimension reduction in classification," in *IEEE Congress on Evolutionary Computation (CEC)*, pp. 2428–2435, 2013.

[168] C. Bae, W.-C. Yeh, Y. Y. Chung, and S.-L. Liu, "Feature selection with intelligent dynamic swarm and rough set," *Expert Systems with Applications*, vol. 37, no. 10, pp. 7026 – 7032, 2010.

[169] S. S. S. Ahmad and W. Pedrycz, "Feature and instance selection via cooperative PSO," in *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 2127–2132, 2011.

[170] L. Cervante, B. Xue, L. Shang, and M. Zhang, "A dimension reduction approach to classification based on particle swarm optimisation and rough set theory," in *25nd Australasian Joint Conference on Artificial Intelligence*, vol. 7691 of *Lecture Notes in Computer Science*, pp. 313–325, Springer, 2012.

[171] M. A. Mandal M, "A graph-theoretic approach for identifying non-redundant and relevant gene markers from microarray data using multiobjective binary PSO," *PLoS ONE*, vol. 9, pp. 1–13, 2014.

[172] L. Cervante, B. Xue, L. Shang, and M. Zhang, "A multi-objective feature selection approach based on binary PSO and rough set theory," in *13th European Conference on Evolutionary Computation in Combinatorial Optimization (EvoCOP)*, vol. 7832 of *Lecture Notes in Computer Science*, pp. 25–36, Springer, 2013.

[173] B. Xue, L. Cervante, L. Shang, W. N. Browne, and M. Zhang, "A multi-objective particle swarm optimisation for filter based feature selection in classification problems," *Connection Science*, vol. 24, no. 2-3, pp. 91–116, 2012.

[174] B. Xue, L. Cervante, L. Shang, W. N. Browne, and M. Zhang, "Binary PSO and rough set theory for feature selection: A multi-objective filter based approach," *International Journal of Computational Intelligence and Applications*, vol. 13, no. 02, pp. 1450009:1–34, 2014.

[175] M. A. Esseghir, G. Goncalves, and Y. Slimani, "Adaptive particle swarm optimizer for feature selection," in *international conference on Intelligent data engineering and automated learning (IDEAL)*, (Berlin, Heidelberg), pp. 226–233, Springer Verlag, 2010.

[176] B. Sahu and D. Mishra, "A novel feature selection algorithm using particle swarm optimization for cancer microarray data," *Procedia Engineering*, vol. 38, pp. 27 – 31, 2012.

[177] H. B. Nguyen, B. Xue, I. Liu, and M. Zhang, "Filter based backward elimination in wrapper based PSO for feature selection in classification," in *IEEE Congress on Evolutionary Computation (CEC)*, pp. 3111 – 3118, 2014.

[178] R. Huang and M. He, "Feature selection using double parallel feedforward neural networks and particle swarm optimization," in *IEEE Congress on Evolutionary Computation (CEC)*, pp. 692–696, 2007.

[179] A. Boubezoul and S. Paris, "Application of global optimization methods to model and feature selection," *Pattern Recognition*, vol. 45, no. 10, pp. 3676 – 3686, 2012.

[180] R. Fdhila, T. Hamdani, and A. Alimi, "Distributed MOPSO with a new population subdivision technique for the feature selection," in *International Symposium on Computational Intelligence and Intelligent Informatics (ISCIII)*, pp. 81 –86, 2011.

[181] R. M. Ramadan and R. F. Abdel Kader, "Face recognition using particle swarm optimization-based selected features," *International Journal of*

Signal Processing, Image Processing and Pattern Recognition, vol. 2, no. 2, pp. 51–65, 2007.

[182] J. Kennedy and R. Eberhart, "A discrete binary version of the particle swarm algorithm," in *IEEE International Conference on Systems, Man, and Cybernetics*, vol. 5, pp. 4104–4108, 1997.

[183] R. Jensen and Q. Shen, "Finding rough set reducts with ant colony optimization," in *Proceedings of the 2003 UK Workshop on Computational Intelligence*, pp. 15–22, 2003.

[184] H. H. Gao, H. H. Yang, and X. Y. Wang, "Ant colony optimization based network intrusion feature selection and detection," in *International Conference on Machine Learning and Cybernetics*, vol. 6, pp. 3871–3875, 2005.

[185] R. K. Sivagaminathan and S. Ramakrishnan, "A hybrid approach for feature subset selection using neural networks and ant colony optimization," *Expert Systems with Applications*, vol. 33, no. 1, pp. 49–60, 2007.

[186] S. Vieira, J. Sousa, and T. Runkler, "Ant colony optimization applied to feature selection in fuzzy classifiers," in *Foundations of Fuzzy Logic and Soft Computing*, vol. 4529 of *Lecture Notes in Computer Science*, pp. 778–788, Heidelberg, 2007.

[187] H. R. Kanan and K. Faez, "An improved feature selection method based on ant colony optimization (ACO) evaluated on face recognition system," *Applied Mathematics and Computation*, vol. 205, no. 2, pp. 716–725, 2008.

[188] N. M. O'Boyle, D. S. Palmer, F. Nigsch, and J. B. Mitchell, "Simultaneous feature selection and parameter optimisation using an artificial ant colony: case study of melting point prediction," *Chemistry Central Journal*, vol. 2, no. 1, pp. 1–15, 2008.

[189] H. Yu, G. Gu, H. Liu, J. Shen, and J. Zhao, "A modified ant colony optimization algorithm for tumor marker gene selection," *Genomics, Proteomics & Bioinformatics*, vol. 7, no. 4, pp. 200–208, 2009.

[190] S. M. Vieira, J. M. Sousa, and T. A. Runkler, "Two cooperative ant colonies for feature selection using fuzzy models," *Expert Systems with Applications*, vol. 37, no. 4, pp. 2714–2723, 2010.

[191] X. Li, W. Zhu, B. Ji, B. Liu, and C. Ma, "Shape feature selection and weed recognition based on image processing and ant colony optimization," *Transactions of the Chinese Society of Agricultural Engineering*, vol. 26, no. 10, pp. 178–182, 2010.

[192] S. Kashef and H. Nezamabadi-pour, "A new feature selection algorithm based on binary ant colony optimization," in *Information and Knowledge Technology (IKT), 2013 5th Conference on*, pp. 50–54, 2013.

[193] S. Kashef and H. Nezamabadi-pour, "An advanced ACO algorithm for feature subset selection," *Neurocomputing*, 2014.

[194] S. Vieira, J. Sousa, and T. Runkler, "Multi-criteria ant feature selection using fuzzy classifiers," in *Swarm Intelligence for Multi-objective Problems in Data Mining*, vol. 242 of *Studies in Computational Intelligence*, pp. 19–36, Heidelberg, 2009.

[195] C.-K. Zhang and H. Hu, "Feature selection using the hybrid of ant colony optimization and mutual information for the forecaster," in *International Conference on Machine Learning and Cybernetics*, vol. 3, pp. 1728–1732, 2005.

[196] R. Jensen, "Performing feature selection with aco," in *Swarm Intelligence in Data Mining*, vol. 34 of *Studies in Computational Intelligence*, pp. 45–73, / Heidelberg, 2006.

[197] L. Ke, Z. Feng, and Z. Ren, "An efficient ant colony optimization approach to attribute reduction in rough set theory," *Pattern Recognition Letters*, vol. 29, no. 9, pp. 1351–1357, 2008.

[198] M. He, "Feature selection based on ant colony optimization and rough set theory," in *International Symposium on Computer Science and Computational Technology*, vol. 1, pp. 247–250, 2008.

[199] S. Zhou, J. ping Zhang, and B.-K. Su, "Feature selection and classification based on ant colony algorithm for hyperspectral remote sensing images," in *2nd International Congress on Image and Signal Processing*, pp. 1–4, 2009.

[200] Y. Chen, D. Miao, and R. Wang, "A rough set approach to feature selection based on ant colony optimization," *Pattern Recognition Letters*, vol. 31, no. 3, pp. 226–233, 2010.

[201] H. Huang, H.-B. Xie, J.-Y. Guo, and H.-J. Chen, "Ant colony optimization-based feature selection method for surface electromyography signals classification," *Computers in Biology and Medicine*, vol. 42, no. 1, pp. 30–38, 2012.

[202] S. Ali and W. Shahzad, "A feature subset selection method based on symmetric uncertainty and ant colony optimization," in *2012 International Conference on Emerging Technologies (ICET)*, pp. 1–6, 2012.

[203] A. Tallón-Ballesteros and J. Riquelme, "Tackling ant colony optimization meta-heuristic as search method in feature subset selection based on correlation or consistency measures," in *Intelligent Data Engineering and Automated Learning (IDEAL)*, vol. 8669 of *Lecture Notes in Computer Science*, pp. 386–393, Springer, 2014.

[204] L. Ke, Z. Feng, Z. Xu, K. Shang, and Y. Wang, "A multiobjective aco algorithm for rough feature selection," in *Second Pacific-Asia Conference on Circuits,Communications and System (PACCS)*, vol. 1, pp. 207–210, 2010.

[205] M. M. Kabir, M. Shahjahan, and K. Murase, "A new hybrid ant colony optimization algorithm for feature selection," *Expert Systems with Applications*, vol. 39, no. 3, pp. 3747–3763, 2012.

[206] M. He, "A rough set based hybrid method to feature selection," in *International Symposium on Knowledge Acquisition and Modeling (KAM)*, pp. 585–588, IEEE, 2008.

[207] B. Xue, W. Fu, and M. Zhang, "Multi-objective feature selection in classification: A differential evolution approach," in *Simulated Evolution and Learning*, vol. 8886 of *Lecture Notes in Computer Science*, pp. 516–528, 2014.

[208] C. M. Wang and Y. F. Huang, "Evolutionary-based feature selection approaches with new criteria for data mining: A case study of credit approval data," *Expert Systems with Applications*, vol. 36, pp. 5900–5908, 2009.

[209] X. Liu, C. Yu, and Z. Cai, "Differential evolution based band selection in hyperspectral data classification," in *Advances in Computation and Intelligence*, vol. 6382 of *Lecture Notes in Computer Science*, pp. 86–94, Heidelberg, 2010.

[210] R. N. Khushaba, A. Al-Ani, and A. Al-Jumaily, "Feature subset selection using differential evolution and a statistical repair mechanism," *Expert Systems with Applications*, vol. 38, no. 9, pp. 11515–11526, 2011.

[211] M. Omidvar, X. Li, Y. Mei, and X. Yao, "Cooperative co-evolution with differential grouping for large scale optimization," *IEEE Transactions on Evolutionary Computation*, vol. 18, no. 3, pp. 378–393, 2014.

[212] Z. Yang, K. Tang, and X. Yao, "Scalability of generalized adaptive differential evolution for large-scale continuous optimization," *Soft Computing*, vol. 15, pp. 2141–2155, 2011.

[213] K. Chakravarty, D. Das, A. Sinha, and A. Konar, "Feature selection by differential evolution algorithm - a case study in personnel identification," in *IEEE Congress on Evolutionary Computation (CEC)*, pp. 892–899, 2013.

[214] C.-S. Yang, L.-Y. Chuang, Y.-J. Chen, and C.-H. Yang, "Feature selection using memetic algorithms," in *3rd International Conference on Convergence and Hybrid Information Technology*, vol. 1, pp. 416–423, 2008.

[215] M. Abedini, M. Kirley, and R. Chiong, "Incorporating feature ranking and evolutionary methods for the classification of high-dimensional dna microarray gene expression data," *Australasian Medical Journal*, vol. 6, no. 5, pp. 272–279, 2013.

[216] A. Al-Ani, A. Alsukker, and R. N. Khushaba, "Feature subset selection using differential evolution and a wheel based search strategy," *Swarm and Evolutionary Computation*, vol. 9, pp. 15–26, 2013.

[217] P. S. Bharathi P T, "Optimal feature subset selection using differential evolution and extreme learning machine," *International Journal of Science and Research (IJSR)*, vol. 3, pp. 1898–1905, 2014.

[218] Z. Li, Z. Shang, B. Qu, and J. Liang, "Feature selection based on manifold-learning with dynamic constraint handling differential evolution," in *IEEE Congress on Evolutionary Computation (CEC)*, pp. 332–337, 2014.

[219] I.-S. Oh, J.-S. Lee, and B.-R. Moon, "Hybrid genetic algorithms for feature selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 11, pp. 1424 –1437, 2004.

[220] O. Alomari and Z. A. Othman, "Bees algorithm for feature selection in network anomaly detection," *Journal of Applied Sciences Research*, vol. 8, no. 3, pp. 1748–1756, 2012.

[221] S. Palanisamy and S. Kanmani, "Artificial bee colony approach for optimizing feature selection," *International Journal of Computer Science Issues (IJCSI)*, vol. 9, no. 3, pp. 432–438, 2012.

[222] G. V. Vahid Chahkandi, Mahdi Yaghoobi, "Feature selection with chaotic hybrid artificial bee colony algorithm based on fuzzy (chabcf)," *Journal of Soft Computing and Applications*, vol. 2013, pp. 1–8, 2013.

[223] "A binary {ABC} algorithm based on advanced similarity scheme for feature selection," *Applied Soft Computing*, vol. 36, p. 334–348, 2015.

[224] K. Shelke, S. Jayaraman, S. Ghosh, and J. Valadi, "Hybrid feature selection and peptide binding affinity prediction using an EDA based algorithm," in *IEEE Congress on Evolutionary Computation (CEC)*, pp. 2384–2389, 2013.

[225] X. Han, X. Chang, L. Quan, X. Xiong, J. Li, Z. Zhang, and Y. Liu, "Feature subset selection by gravitational search algorithm optimization," *Information Sciences*, vol. 281, pp. 128–146, 2014.

[226] H. Zhang and G. Sun, "Feature selection using tabu search method," *Pattern Recognition*, vol. 35, pp. 701–711, 2002.
[227] Z. Zhu, S. Jia, and Z. Ji, "Towards a memetic feature selection paradigm [application notes]," *IEEE Computational Intelligence Magazine*, vol. 5, no. 2, pp. 41–53, 2010.
[228] Z. Zhu and Y.-S. Ong, "Memetic algorithms for feature selection on microarray data," in *Advances in Neural Networks (ISNN)*, vol. 4491 of *Lecture Notes in Computer Science*, pp. 1327–1335, Heidelberg, 2007.
[229] J. Huang, Y. Cai, and X. Xu, "A hybrid genetic algorithm for feature selection wrapper based on mutual information," *Pattern Recognition Letters*, vol. 28, no. 13, pp. 1825 – 1844, 2007.
[230] M. Esseghir, G. Goncalves, and Y. Slimani, "Memetic feature selection: Benchmarking hybridization schemata," in *Hybrid Artificial Intelligence Systems*, vol. 6076 of *Lecture Notes in Computer Science*, pp. 351–358, Heidelberg, 2010.
[231] Y. Wen and H. Xu, "A cooperative coevolution-based pittsburgh learning classifier system embedded with memetic feature selection," in *IEEE Congress on Evolutionary Computation*, pp. 2415–2422, 2011.
[232] M. Schiezaro and H. Pedrini, "Data feature selection based on artificial bee colony algorithm," *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, pp. 1–8, 2013.
[233] J. Jelonek, K. Krawiec, and R. Sowiski, "Rough set reduction of attributes and their domains for neural networks," *Computational Intelligence*, vol. 11, no. 2, pp. 339–347, 1995.
[234] B. Xue, *Particle Swarm Optimisation for Feature Selection*. PhD thesis, Victoria University of Wellington, Wellington, New Zealand, 2014.
[235] D. Kumar, D. Kumar, and C. Rai, "Memetic algorithms for feature selection in face recognition," in *International Conference on Hybrid Intelligent Systems*, pp. 931–934, 2008.
[236] E. Alpaydin, *Introduction to machine learning*. The MIT Press, 2004.
[237] W. Dong, T. Chen, P. Tino, and X. Yao, "Scaling up estimation of distribution algorithms for continuous optimization," *IEEE Transactions on Evolutionary Computation*, vol. 17, no. 6, pp. 797–822, 2013.
[238] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *Journal of Machine Learning Research*, vol. 5, pp. 1205 – 1224, 2004.
[239] E. Amaldi and V. Kann, "On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems," *Theoretical Computer Science*, vol. 209, no. 1–2, pp. 237–260, 1998.
[240] M. N. Le, Y. S. Ong, S. Menzel, Y. Jin, and B. Sendhoff, "Evolution by adapting surrogates," *Evolutionary Computation*, vol. 21, no. 2, pp. 313–340, 2013.
[241] P. Estevez, M. Tesmer, C. Perez, and J. Zurada, "Normalized mutual information feature selection," *IEEE Transactions on Neural Networks*, vol. 20, no. 2, pp. 189–201, 2009.
[242] A. Jakulin and I. Bratko, "Testing the significance of attribute interactions," in *the 21 International Conference on Machine Learning (ICML)*, pp. 52–59, ACM, 2004.
[243] Y. Zhai, M. Tan, I. Tsang, and Y. S. Ong, "Discovering support and affiliated features from very high dimensions," in *Proceedings of the 29th International Conference on Machine Learning (ICML)*.
[244] C. A. Coello Coello, "Evolutionary multi-objective optimization: a historical view of the field," *IEEE Computational Intelligence Magazine*, vol. 1, no. 1, pp. 28–36, 2006.
[245] V. Khare, X. Yao, and K. Deb, "Performance scaling of multi-objective evolutionary algorithms," in *Evolutionary Multi-Criterion Optimization*, pp. 376–390, Springer, 2003.
[246] K. Praditwong and X. Yao, "How well do multi-objective evolutionary algorithms scale to large problems," in *IEEE Congress on Evolutionary Computation (CEC)*, pp. 3959–3966, 2007.
[247] K. Neshatian, *Feature Manipulation with Genetic Programming*. PhD thesis, Victoria University of Wellington, Wellington, New Zealand, 2010.
[248] C. Ambroise and G. J. McLachlan, "Selection bias in gene extraction on the basis of microarray gene-expression data," *Proceedings of the National Academy of Sciences*, vol. 99, no. 10, pp. 6562–6566, 2002.
[249] S. K. Singhi and H. Liu, "Feature subset selection bias for classification learning," in *23rd International Conference on Machine Learning (ICML)*, pp. 849–856, ACM, 2006.
[250] U. M. Braga-Neto and E. R. Dougherty, "Is cross-validation valid for small-sample microarray classification ?," *Bioinformatics*, vol. 20, no. 3, pp. 374–380, 2004.
[251] J. Tang and H. Liu, "Coselect: Feature selection with instance selection for social media data," in *SIAM International Conference on Data Mining (SDM)*, pp. 695–694, 2013.

Bing Xue (M'10) received her PhD degree in 2014 at Victoria University of Wellington, New Zealand. Since May 2015, she has been working as a Lecturer at Victoria University of Wellington. She is with the Evolutionary Computation Research Group at VUW, and her research focuses mainly on evolutionary computation, machine learning and data mining, particularly, evolutionary computation for feature selection, feature construction, dimension reduction, symbolic regression, transfer learning, domain adaptation, image analysis, multi-objective optimisation, bioinformatics and big data. Dr Xue is currently leading the strategic research direction on evolutionary feature selection and construction in Evolutionary Computation Research Group at VUW, and has been organsing special sessions and issues on evolutionary computation for feature selection and construction. Dr Xue is a member of IEEE CIS Evolutionary Computation Technical Committee, and she is also the Chair of IEEE CIS Task Force on Evolutionary Computation for Feature Selection and Construction.

Mengjie Zhang (SM'10) is currently Professor of Computer Science at Victoria University of Wellington, where he heads the interdisciplinary Evolutionary Computation Research Group. His research is mainly focused on evolutionary computation, particularly genetic programming, particle swarm optimisation, multi-objective optimisation and learning classifier systems with application areas of classification with unbalanced data, feature selection, computer vision and image processing, job shop scheduling, transfer learning, and bioinformatics. He is also interested in data mining, machine learning, and web information extraction. He has published over 350 academic papers in refereed international journals and conferences in these areas. Since 2007, he has been listed as one of the top ten world genetic programming researchers by the GP bibliography. Prof Zhang is the Chair of the IEEE CIS Evolutionary Computation Technical Committee, and a member of IEEE CIS Intelligent System Applications Technical Committee.

Will N. Browne is an Associate Professor with the Evolutionary Computation Research Group, School of Engineering and Computer Science, Victoria University of Wellington, New Zealand after serving eight years at the Department of Cybernetics, University of Reading, U.K. His main research interest is applied cognitive systems, i.e. developing both virtual and real-world systems that can perceive, represent, reason, learn and effect actions that address complex problems. Specific interests include Learning Classifier Systems and other branches of Evolutionary Computation/Swarm Intelligence, Cognitive Robotics, and Modern Heuristics for industrial applications. Current research projects include evolutionary computation vision and transfer of learnt, abstracted knowledge/functionality. He has published over 100 academic papers in books, refereed international journals, and conferences.

Xin Yao is a Professor of Computer Science and the Director of CERCIA (the Centre of Excellence for Research in Computational Intelligence and Applications) at the University of Birmingham, UK. He is an IEEE Fellow and the President (2014-15) of IEEE Computational Intelligence Society (CIS). His major research interests include evolutionary computation and ensemble learning. His research won the 2001 IEEE Donald G. Fink Prize Paper Award, 2010 and 2015 IEEE Transactions on Evolutionary Computation Outstanding Paper Award, 2010 BT Gordon Radley Award for Best Author of Innovation (Finalist), 2011 IEEE Transactions on Neural Networks Outstanding Paper Award, and many other best paper awards. He received the prestigious Royal Society Wolfson Research Merit Award in 2012 and the IEEE CIS Evolutionary Computation Pioneer Award in 2013. He was the Editor-in-Chief (2003-08) of IEEE Transactions on Evolutionary Computation. He is a co-inventor of an international patent on "Exploiting ensemble diversity for automatic feature extraction".