A Grid-dominance based Multi-objective Algorithm for Feature Selection in Classification

Peng Wang, Bing Xue, Mengjie Zhang
School of Engineering and Computer Science
Victoria University of Wellington, PO Box 600
Wellington 6140, New Zealand
{wangpeng,bing.xue,mengjie.zhang}@ecs.vuw.ac.nz

Jing Liang
School of Electrical Engineering
Zhengzhou University
Zhengzhou, China
liangjing@zzu.edu.cn

***Abstract*—Feature selection aims to select a small subset of relevant features while maintaining or even improving the classification performance over using all features. Feature selection can be considered as a multi-objective problem, i.e., minimizing the number of selected features and maximizing the classification accuracy (minimizing the classification error) simultaneously. Most evolutionary multi-objective algorithms encounter difficulties when handling a feature selection task due to the discrete search space, although they perform well on continuous/numeric optimization problems. This paper proposes a grid-dominance based multi-objective evolutionary algorithm to address feature selection. The aim is to explore the potential of the grid-dominance method to strengthen the selection pressure toward the optimal direction while maintaining an extensive distribution among the objective values of feature subsets. To increase the population diversity, a subset filtration mechanism is proposed. The performance of the proposed two algorithms is tested on fourteen datasets of varying difficulty. With the proposed methods, the performance metrics, hypervolume and inverted generational distance have been significantly improved compared with other commonly used multi-objective algorithms, and the population diversity has also been increased.**

I. Introduction

The advance in data collection promotes the fast-growing of high-dimensional data (i.e., the number of features). However, only a portion of features are relevant to predict the class label in a classification task in the usual case [1]. Meanwhile, the involvement of irrelevant and/or redundant features may deteriorate the classification accuracy of a learning algorithm [2], since irrelevant or redundant features may obscure useful information from relevant or discriminating features. By eliminating irrelevant and redundant features, feature selection has been applied to help different learning algorithms to obtain better classification performance in a shorter time [3].

However, feature selection is not an easy task due to the following reasons. First, exhaustively searching for optimal feature subsets (i.e., solutions) in 2^n possible solutions with n features is difficult on a high-dimensional dataset [4]. Second, feature selection can be taken as a bi-objective problem where the two main objectives, maximizing the classification accuracy (minimizing the classification error) and minimizing the number of selected features, are potentially in conflict [5]. Another challenge of feature selection is the complex interactions among features. A feature that is individually irrelevant to the class may be highly useful when it is complementary to other features. On the other hand, an individually relevant feature may become redundant or even noisy when working with other features.

Feature selection approaches mainly belong to four main categories: filter, wrapper, hybrid, and embedded methods [6]. A filter method usually uses certain statistical measures, such as information theory and correlation measures, to evaluate the goodness of a feature subset in a way independent of any learning algorithm. In contrast, in wrapper methods, feature subsets are evaluated by a specific classification algorithm such as K-nearest neighbour (KNN) [7]. Hybrid methods typically combine wrapper with filter methods to select useful features, whereas embedded methods integrate learning and feature selection into a single process. Generally, filter methods often execute faster and have lower computational complexity while having lower classification accuracy than wrapper methods. Therefore, the wrapper method is utilized in this paper to evaluate the candidate feature subsets.

As a population based search technique, evolutionary computation (EC) has been the leading method in evolving a set of trade-off solutions for multi-objective problems, including feature selection [2]. Many evolutionary multi-objective optimization (EMO) algorithms have been proposed, including non-dominated sorting genetic algorithm II (NSGA-II) [8], strength Pareto evolutionary algorithm 2 (SPEA2) [9], and multi-objective evolutionary algorithm based on decomposition (MOEA/D) [10]. In EMO methods, the concept of Pareto-dominance [8] is needed to distinguish solutions during the environmental selection process. Some EMO methods choose parents based on the Pareto-dominance relationship in an elitist manner to generate off-springs [11].

However, the Pareto-dominance relationship in most EMO based feature selection algorithms needs to be improved. Yue et al. [12] pointed out that when using EMO methods to deal with a problem with discontinuous search space (e.g., sparse reconstruction or feature selection), some optimal solutions may be lost if the dominated solutions are completely disregarded during evolution. The unexplored parts of the search space may well contain effective solutions (e.g., S_9 in Fig.

978-1-7281-8393-0/21/$31.00 ©2021 IEEE

1) that might be of great interest to a user. Another obstacle of using many EMO methods to deal with feature selection is the frequently occurred duplications [13] (absence of the diversity among individuals), i.e., multiple vectors result in the same feature subset, which may lead to a local optimum. Considering the worst case where all the individuals map to the same feature subset, the population has no diversity at all both in the solution space and the objective space, which easily makes the algorithm converge prematurely.

A feasible way to search for a diverse set of feature subsets is to use a relaxed form of the Pareto-dominance that allows a user to regulate the granularity of the approximation of fitness values of feature subsets, e.g., grid-dominance [14]. In grid-dominance, the mutual relationship of individuals is determined in a grid environment, which provides a simple yet effective tool to quantify each individual's performance [15]. Grid-dominance has already shown its advantages in several EMO methods [14]–[16]. Furthermore, to increase the population diversity, the subset filtration mechanism based on the concept of confidence level/rate [17] is proposed to filter and pick feature subsets from the duplicated ones. By exploring the utilities of grid-dominance and the proposed subset filtration mechanism, a feature selection algorithm based on NSGAII is proposed, named GF-NSGAII. In the EMO community, NSGAII is the representative one, which is suitable to test the performance of the proposed mechanism.

A. Goals

The overall goal of this work is to design a grid-dominance based multi-objective feature selection algorithm to improve the classification performance and reduce the number of selected features. To achieve this goal, a subset filtration mechanism is specifically designed for feature selection, which is then applied to NSGAII to test its performance (termed F-NSGAII). Moreover, the concept of grid-dominance is applied to the proposed GF-NSGAII algorithm. The proposed two algorithms are compared against four other EMO based algorithms on fourteen datasets. Specifically, we will investigate:

1) whether integrating the proposed subset filtration mechanism can improve the population diversity, and
2) whether the concept of grid-dominance can help the algorithm to obtain better feature subsets compared with the Pareto-dominance, and
3) whether using the proposed mechanism significantly increases the running time.

II. Related Work

A. Multi-objective Optimization and the Pareto-dominance

Under the framework of multi-objective optimization, two or more conflicting objectives are usually optimized simultaneously. An o-objective minimization problem can be stated as follows:

$$\begin{aligned} \min\ & \vec{f}(\vec{x}) = (f_1(\vec{x}), f_2(\vec{x}), \ldots, f_o(\vec{x})) \\ \text{subject to } & g_i(\vec{x}) \leq 0 \quad i = 1, 2, \ldots, k \\ & h_j(\vec{x}) = 0 \quad j = 1, 2, \ldots, p \end{aligned} \tag{1}$$

where $\vec{x}$ represents the decision variable vector, and $\vec{f}(\vec{x})$ is a vector with o objective functions; $f_o(\vec{x})$ is the o-th objective. Meanwhile, $g_i(\vec{x})$ and $h_j(\vec{x})$ mean the constraint functions of the problem.

According to the definition of dominance in [18], the goodness of a solution in multi-objective optimization is based on the Pareto-dominance relationship. A feasible solution $\vec{r}$ is better than another feasible solution $\vec{t}$ if:

$$\forall\, i : f_i(\vec{r}) \leq f_i(\vec{t}) \ \text{ and } \ \exists\, j : f_j(\vec{r}) < f_j(\vec{t}) \tag{2}$$

It can be said that $\vec{r}$ dominates $\vec{t}$ ($i, j \in \{1, 2, \ldots, o\}$). A solution $\vec{x}^*$ is a Pareto-optimal solution, on the condition that it is not dominated by any other feasible solution. The set of all the Pareto-optimal solutions forms Pareto front (PF) in the objective space. An EMO method aims to evolve a well-distributed PF [5], [19].

B. Grid-dominance and Fitness Assignment

The concept of grid-dominance was utilized in [14], where multiple grids are utilized as a frame to determine the location/coordinate of individuals in the objective space. The grid-dominance is similar to the Pareto-dominance, and the grid coordinates of individuals in grid-dominance play the same role as their actual objective values in Pareto-dominance [14]. Fig. 1 illustrates individuals in a grid environment in the objective space. For solutions S_1, S_2, and S_3 in the figure, their grid coordinates are $(0, 3)$, $(2, 3)$, and $(1, 2)$, respectively. For solutions S_6-S_9, their grid coordinates are the same: $(3, 0)$.

Some solutions have the same grid coordinates (e.g., S_6 to S_8 in Fig. 1), and thus a high chance of being eliminated or preserved simultaneously. Following the idea in [14], a fitness assignment mechanism is introduced to address this issue. Three grid based criteria, i.e., grid ranking (GR), grid crowding distance (GCD), and grid coordinate point distance ($GCPD$), are taken into account to assign the fitness of individuals.

$$GR(\vec{x}) = \sum_{i=1}^{o} G_i(\vec{x}) \tag{3}$$

where $G_i(\vec{x})$ means the grid coordinate of individual $\vec{x}$ in the i-th objective, and o is the number of objectives.

$$GD(\vec{x}, \vec{y}) = \sum_{i=1}^{o} |G_i(\vec{x}) - G_i(\vec{y})| \tag{4}$$

where $GD(\vec{x}, \vec{y})$ stands for the grid difference (GD) between $\vec{x}$ and $\vec{y}$.

$$GCD(\vec{x}) = (\sum_{\vec{y} \in N(\vec{x})} o - GD(\vec{x}, \vec{y})) \tag{5}$$

where $N(\vec{x})$ represents the set of neighbors of $\vec{x}$. A solution $\vec{y}$ is taken as a neighbor of a solution $\vec{x}$ if $GD(\vec{x}, \vec{y}) < o$.

$$GCPD(\vec{x}) = \sqrt{\sum_{i=1}^{o} ((F_i(\vec{x}) - (lb_i + G_i(\vec{x}) * d_i))/d_i)^2} \tag{6}$$

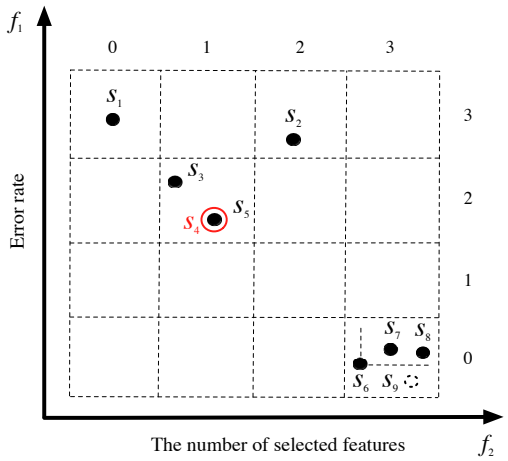

Fig. 1: Illustration of solutions in a grid environment in a bi-objective space. The f_1 and f_2 are the classification error rate and the number of selected features, respectively. In this example, the number of the divisions of the objective space is four (i.e., $div = 4$). Suppose that these points represent the distribution of solutions for a feature selection task in a certain generation. Solutions/subsets S_4 and S_5 sit at the same point in the objective space. S_9 is a hypothetical solution, which does not exist in this generation.

where $G_i(\vec{x})$ and $F_i(\vec{x})$ mean the grid coordinate and actual objective values of individual $\vec{x}$ in the i-th objective, respectively. Meanwhile, lb_i and ub_i represent the lower and upper boundaries of the grid, respectively. The width of a hyperbox for the i-th objective is denoted as d_i, $d_i = (ub_i - lb_i/div)$; div is the number of the divisions of the objective space in each dimension (e.g., in Fig. 1, $div = 4$).

GR and $GCPD$ are concerned with the population convergence while GCD is used to measure the population diversity in the objective space. GR is regarded as the primary criterion, and GCD is the secondary one, activated when the GR value of individuals is the same. When the first two criteria fail to discriminate individuals, the third, $GCPD$ is used to break a tie. For these three indicators, the smaller the better.

III. Proposed Feature Selection Method

A. Representation and Objective Functions

The representation of an individual in NSGAII is a vector of real numbers (between 0 and 1). The dimensionality of each vector is equal to the number of original features in one dataset, where a pre-defined threshold θ is employed to determine whether the feature is chosen or not. If the corresponding value in the vector (i.e., position entry) is greater than the threshold θ, the feature will be chosen, and otherwise not chosen. During evaluation, the classification performance is obtained based on the selected features.

The objective functions are shown in Eq. (7), which are to minimize the classification error rate and the number of selected features simultaneously.

$$\min \begin{cases} f_1 = ER = \frac{FP+FN}{TP+TN+FP+FN} \\ f_2 = FR = \frac{\#Selected\ Features}{\#\text{Original Features}} \end{cases} \tag{7}$$

where ER means error rate and FR is the rate of the number of the selected features over the total number of original features. Meanwhile, FP, FN, TP, and TN are the false positives, false negatives, true positives, and true negatives, respectively.

B. General Procedure of the Proposed Algorithm

Most existing EMO based feature selection algorithms, including NSGAII, use the Pareto-dominance during evolution to choose subsets for the next generation. Take Fig. 1 as an example to show the situation. Suppose that the algorithm needs to pick five individuals from S_1 to S_8 to enter into the next generation (S_9 is a hypothetical solution). According to the definition of the Pareto-dominance, S_2 is dominated by S_3-S_5, and S_7-S_8 are dominated by S_6. Therefore, S_1 and S_3-S_6 will be chosen. However, excluding S_7 and S_8 may lead to the algorithm failing to find the solution S_9, but S_9 is generally preferred since it has the lowest classification error among the whole population. Another issue is that some duplicated solutions (i.e., S_4 and S_5) exist in the chosen feature subsets, which reduces the population diversity in the solution space.

To remove the duplicated solutions, a subset filtration mechanism is proposed and applied to the proposed GF-NSGAII algorithm. During the environmental selection process, only one solution will be selected from the multiple duplicated solutions (e.g., S_4 and S_5) to enter into the next generation. The details of the proposed subset filtration mechanism will be shown in the next subsection, which aims to improve the population diversity in the solution space. Furthermore, the grid-dominance is applied to GF-NSGAII, where the solution S_7 or S_8 can be kept during evolution, since S_6-S_8 all sit in the first front based on grid-dominance. This increases the probability of GF-NSGAII to find the solution S_9.

The flowchart of GF-NSGAII is given in Fig. 2. The main processes of GF-NSGAII are similar to NSGAII. After the generation of off-springs, the parents and off-springs are combined. After performing the subset filtration mechanism, if the preserved number of solutions is equal to the population size, there is no need to perform the fitness assignment (shown in Section II). Otherwise, the GR, GCD, and $GCPD$ values of all the preserved individuals are calculated based on Eqs. (3) to (6). Next, a certain number (i.e., pop size) of solutions are selected based on their fitness values to form the population in the next generation. When the stopping criterion is met, the algorithm will output the non-dominated solutions.

C. Subset Filtration Mechanism

The more diverse the population, the better the user can understand the possible solutions [20]. When multiple individuals are resulting in the same feature subset (i.e., multiple duplicated solutions), the individual which has the largest sum of the confident rate will be kept among the multiple duplicated solutions, and the others will be removed.

Nguyen et al. [17] presented a method to calculate the degree of confidence in the decision of feature selection. The higher the confidence, the less likely the decision is changed.

$$CR(i) = \begin{cases} \frac{f_i - \theta}{1 - \theta} & \text{if } f_i > \theta \\ \frac{\theta - f_i}{\theta} & \text{if } f_i \leq \theta \end{cases} \tag{8}$$

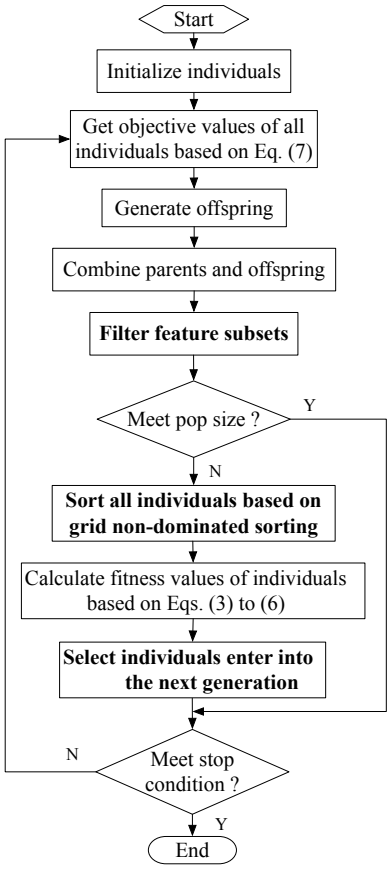

Fig. 2: The flowchart of the proposed GF-NSGAII algorithm.

Algorithm 1: Subset filtration mechanism

Input: S: A set of feature subsets
Output: US: A set of diverse feature subsets

```
1  begin
2      if All selected features are different between each other in S
         then
3          US ← S
4      else
5          // S_dup: All duplicated feature subsets in S
6          // S_dup = S_dup_1 ∪ S_dup_2 ∪ ... ∪ S_dup_p
7          // US ← S\S_dup
8          for i = 1 to p do
9              // Calculate SCR values of all individuals in S_dup_i
                 based on Eq. (9)
10             // index = find_maximum_index[SCR]
11             US ← S_dup_i(index)
12         end
13     end
14 end
```

TABLE I: The information of datasets

Number	Dataset	# Features	# Classes	# Instances
1	Wine	13	3	178
2	Zoo	16	7	101
3	SPECT	22	2	267
4	WBCD	30	2	569
5	Ionosphere	34	2	351
6	Sonar	60	2	208
7	Movementlibras	90	15	360
8	Hillvally	100	2	606
9	Musk1	166	2	476
10	Multiple(pix)	240	10	2000
11	Arrhythmia	279	16	452
12	SRBCT	2308	4	83
13	Leukemia	5147	2	72
14	DLBCL	7050	2	77

where $CR(i)$ represents the confidence rate on the i-th feature, and f_i is the position entry in the i-th dimension ($i \in \{1, 2, \ldots, N\}$, N is the number of the original features).

$$SCR(\vec{x}) = \sum_{i=1}^{N} CR(i) \tag{9}$$

where $SCR(\vec{x})$ means the sum of the confidence rate of the individual $\vec{x}$. The larger the SCR, the more confidence to choose the corresponding feature(s).

Based on the SCR values of the duplicated solutions, the proposed subset filtration mechanism is shown in Algorithm 1. The first step is to check whether there are duplicated solutions in S, and S is a combination of current parents and off-springs. If the selected feature subsets by all solutions are different from each other, the proposed subset filtration mechanism will stop and output S (Lines 2-3). Otherwise, all feature subsets in S will be screened out and grouped (Lines 5-6). The unique feature subsets in S will be stored into US (Line 7). The next step is to pick solution(s) from one or multiple group(s) of the duplicated solutions. In each group, only the solution that has the largest SCR value will be put into US (Lines 8-11).

The proposed subset filtration mechanism simultaneously considers the unique and the duplicated feature subsets. It can increase the diversity of the population in the solution space since the proposed mechanism is to find and keep potential good and different feature subsets during evolution.

IV. Experiment Design

A. Baseline EMO Methods

To examine the performance of the proposed mechanisms on feature selection, two variant algorithms of NSGAII [18] are proposed. The detailed information can be seen in the following:

* NSGAII with the proposed subset filtration mechanism is named F-NSGAII. The only difference between F-NSGAII and NSGAII is that the subset filtration mechanism is performed in F-NSGAII before the non-dominated sorting.
* The proposed GF-NSGAII algorithm utilizes both subset filtration mechanism and grid-dominance. The procedure of GF-NSGAII is shown in Fig. 2.

The proposed algorithms, F-NSGAII and GF-NSGAII, are compared with four well-known multi-objective algorithms: NSGAII [8], GDE3 [21], SPEA2 [9], and MOEA/D [10].

B. Datasets and Parameter Settings

In this work, six EMO based feature selection algorithms are compared on fourteen different datasets of varying difficulty from the UCI machine learning repository [22]. The selected

datasets have different numbers of classes (2-16), features (13-7050), and instances (72-2000). The information of datasets is shown in Table I.

The grid-dominance in GF-NSGAII has one main parameter: div-the number of the divisions of the objective space in each dimension. The value of div is set to 15 in our experiments. The reason is that the experimental results in [14] showed that a division value around 9 may be reliable on an unknown optimization problem, and a slightly larger div is recommended if the problem is hard to address. For feature selection, the information about the optimal feature subsets is unknown, and different datasets may have different levels of difficulty. Additionally, the settings of GDE3, SPEA2, MOEA/D, and NSGA-II follow the recommended setting from their original papers, which are default settings in the JMetalPy package [23].

Because of the large difference in the search space of different datasets, we set the population size to the number of original features of the target dataset but restricted it to 300 as suggested in [24] to avoid high computational costs. The maximum number of iterations is set to 100, and the threshold θ is 0.6 so as an algorithm begins with a slightly small number of selected features [5].

Each dataset is randomly separated into a training set and a test set with the proportions of 70% and 30%, respectively. Each algorithm has 30 independent runs with 30 different seeds on each training set [19], [25]. In this work, KNN is used as the classifier. Meanwhile, to avoid feature selection bias, KNN with 10-fold cross-validation [2] on the training is used to calculate the classification error rate in Eq. (7). K is set to 5 which is recommended by [5].

C. Performance Indicators

Two indicators are adopted to show the classification performance of different algorithms, which are hypervolume indicator (HV) [26] and inverted generational distance indicator (IGD) [27]. For calculating HV, a reference point is required, which is set to (1, 1). The true PF is unknown for a feature selection task, but it is required when calculating IGD. Therefore, the true PF is approximated by the subsets obtained from the union of all solutions generated by all 30 independent runs of all the six algorithms. The larger the HV or the smaller the IGD, the better the algorithm. Meanwhile, a significance test, the Wilcoxon test with a significance level of 0.05, is used to compare the classification performance of different algorithms.

V. Results

Tables II and III show the average HV and IGD of the six algorithms on the test sets. The two symbols besides the average values of the benchmark algorithms show significant test results compared with the two proposed algorithms, F-NSGAII and GF-NSGAII, respectively. The signs "↑", "↓", "o" represent the corresponding benchmark method is significantly better than, worse than or has no significant difference from F-NSGAII (GF-NSGAII), respectively. The single sign in F-NSGAII's column shows the comparison result with GF-NSGAII. The last row presents the sum of "↑", "o" and "↓" between a pair of algorithms and the rankings according to the Friedman test. In Tables II and III, the best result on the same dataset is highlighted in bold.

Fig. 3 shows the number of unique feature subsets along with the number of generations increases during evolution. To illustrate the impact of the proposed mechanism on the number of unique feature subsets, three algorithms (i.e., NSGAII, F-NSGAII, and GF-NSGAII) on four datasets (namely Wine, Sonar, Multiple, and DLBCL) are selected for comparison. The remaining datasets show similar patterns.

The obtained training and test PFs from the six algorithms are presented in Figs. 4 and 5. Specifically, 30 sets of feature subsets obtained from the 30 runs of each algorithm are combined into a union, in which only non-dominated solutions are retained and shown. Above each figure, the two numbers in bracket mean the total number of the original features and the training or test classification error using all features. The horizontal and vertical axes stand for FR and ER, respectively. Four datasets with a different number of features, i.e., the SPECT, Hillvally, Arrhythmia, and Leukemia datasets, are chosen as representatives due to page limit. The patterns are similar on the other datasets.

A. F-NSGAII vs Benchmark EMO Methods

The significant improvement of F-NSGAII is a result of improvement in the diversity of feature subsets, which can be seen in Fig. 3. The number of unique feature subsets has significantly increased compared with the standard NSGAII algorithm. Furthermore, on the test sets, as shown in Tables II and III, F-NSGAII achieves significantly better HV and IGD results than those of the benchmark algorithms (i.e., GDE3, SPEA2, MOEA/D, and NSGAII) on at least four out of the 14 datasets. Among the five algorithms (GDE3, SPEA2, MOEA/D, NSGAII, and F-NSGAII), F-NSGAII obtains the best HV and IGD values on medium datasets especially on the Movement, Hillvally, Musk1, and Multiple datasets. F-NSGAII shows similar performance with GDE3 and SPEA2, but better performance than MOEA/D and NSGAII on small datasets. However, SPEA2 has better HV and IGD performance on the two large datasets (namely SRBCT and DLBCL). Figs. 4 and 5 show that the fronts evolved by F-NSGAII are usually more diverse than the ones by the four benchmark algorithms. On the Leukemia dataset, the feature subsets obtained by F-NSGAII have a lower test classification error rate, although the feature subsets from SPEA2 have a smaller number of features. The overall ranking of F-NSGAII is 1 among the five algorithms (GDE3, SPEA2, MOEA/D, NSGAII, and F-NSGAII).

Noted that F-NSGAII employs the proposed subset filtration mechanism. The results show that using the proposed mechanism increases the population diversity and therefore helps the algorithm to achieve better HV and IGD results.

TABLE II: The results of HV on the test sets

Dataset	GDE3	SPEA2	MOEA/D	NSGAII	F-NSGAII	GF-NSGAII
Wine	0.847 ± 0.014(o ↓)	0.841 ± 0.017(o ↓)	0.828 ± 0.041(↓ ↓)	0.840 ± 0.027(o ↓)	0.846 ± 0.015(↓)	$\mathbf{0.876 \pm 0.015}$
Zoo	0.830 ± 0.013(o ↓)	0.820 ± 0.018(↓ ↓)	0.811 ± 0.039(↓ ↓)	0.816 ± 0.029(↓ ↓)	0.831 ± 0.010(↓)	$\mathbf{0.842 \pm 0.017}$
SPECT	0.765 ± 0.002(o o)	0.764 ± 0.008(o o)	0.764 ± 0.017(o o)	0.744 ± 0.024(↓ ↓)	0.765 ± 0.013(o)	$\mathbf{0.767 \pm 0.008}$
WBCD	$\mathbf{0.916 \pm 0.003}$(↑ o)	0.913 ± 0.009(o o)	0.896 ± 0.022(↓ ↓)	0.903 ± 0.016(↓ ↓)	0.913 ± 0.008(o)	0.913 ± 0.010
Ionosphere	$\mathbf{0.902 \pm 0.017}$(↑ o)	0.884 ± 0.021(o ↓)	0.851 ± 0.036(↓ ↓)	0.864 ± 0.028(↓ ↓)	0.888 ± 0.022(↓)	0.901 ± 0.016
Sonar	0.843 ± 0.035(o ↓)	0.832 ± 0.038(o ↓)	0.815 ± 0.035(↓ ↓)	0.821 ± 0.033(o ↓)	0.841 ± 0.036(↓)	$\mathbf{0.863 \pm 0.016}$
Movement	0.778 ± 0.013(o o)	0.772 ± 0.018(o o)	0.765 ± 0.023(↓ o)	0.766 ± 0.023(↓ o)	$\mathbf{0.779 \pm 0.017}$(o)	0.772 ± 0.019
Hillvally	0.598 ± 0.007(o ↓)	0.595 ± 0.011(↓ ↓)	0.590 ± 0.013(↓ ↓)	0.585 ± 0.014(↓ ↓)	0.601 ± 0.008(↓)	$\mathbf{0.608 \pm 0.009}$
Musk1	0.929 ± 0.009(↓ ↓)	0.965 ± 0.010(↓ o)	0.963 ± 0.008(↓ o)	0.946 ± 0.011(↓ ↓)	$\mathbf{0.972 \pm 0.006}$(↑)	0.963 ± 0.005
Multiple	0.871 ± 0.008(↓ ↓)	0.937 ± 0.007(↓ ↑)	0.936 ± 0.007(↓ ↑)	0.884 ± 0.017(↓ ↓)	$\mathbf{0.943 \pm 0.006}$(↑)	0.902 ± 0.014
Arrhythmia	0.631 ± 0.013(↓ ↓)	$\mathbf{0.691 \pm 0.020}$(o o)	0.659 ± 0.029(↓ ↓)	0.678 ± 0.018(o ↓)	0.682 ± 0.021(o)	0.690 ± 0.023
SRBCT	0.746 ± 0.010(↓ ↓)	$\mathbf{0.926 \pm 0.019}$(↑ ↑)	0.834 ± 0.008(↓ ↓)	0.895 ± 0.024(↓ o)	0.911 ± 0.015(↑)	0.900 ± 0.007
Leukemia	0.652 ± 0.024(↓ ↓)	0.782 ± 0.048(o ↓)	0.703 ± 0.024(↓ ↓)	0.754 ± 0.049(o ↓)	0.772 ± 0.039(↓)	$\mathbf{0.816 \pm 0.021}$
DLBCL	0.700 ± 0.008(↓ ↓)	$\mathbf{0.834 \pm 0.018}$(↑ ↑)	0.742 ± 0.004(↓ ↓)	0.779 ± 0.019(↓ ↓)	0.803 ± 0.015(↑)	0.796 ± 0.004
Ranking	(2/6/6) (0/4/10) 3.75	(2/8/4) (3/5/6) 2.80	(0/1/13) (1/3/10) 5.18	(0/4/10) (0/2/12) 4.86	(4/4/6) 2.25	2.14

TABLE III: The results of IGD on the test sets

Dataset	GDE3	SPEA2	MOEA/D	NSGAII	F-NSGAII	GF-NSGAII
Wine	0.072 ± 0.010(o ↓)	0.076 ± 0.009(o ↓)	0.081 ± 0.024(↓ ↓)	0.077 ± 0.015(o ↓)	0.073 ± 0.010(↓)	$\mathbf{0.047 \pm 0.014}$
Zoo	0.059 ± 0.016(o o)	0.070 ± 0.016(↓ ↓)	0.082 ± 0.029(↓ ↓)	0.072 ± 0.021(↓ ↓)	0.062 ± 0.012(↓)	$\mathbf{0.053 \pm 0.014}$
SPECT	0.044 ± 0.008(o ↓)	0.047 ± 0.016(o ↓)	0.037 ± 0.017(o ↓)	0.048 ± 0.014(↓ ↓)	0.039 ± 0.014(↓)	$\mathbf{0.024 \pm 0.007}$
WBCD	0.020 ± 0.002(↑ ↓)	0.021 ± 0.006(o ↓)	0.031 ± 0.014(↓ ↓)	0.025 ± 0.010(↓ ↓)	0.021 ± 0.003(↓)	$\mathbf{0.014 \pm 0.007}$
Ionosphere	0.031 ± 0.008(↑ ↓)	0.038 ± 0.009(o ↓)	0.061 ± 0.026(↓ ↓)	0.046 ± 0.013(↓ ↓)	0.037 ± 0.012(↓)	$\mathbf{0.020 \pm 0.013}$
Sonar	0.059 ± 0.022(o ↓)	0.066 ± 0.029(o ↓)	0.077 ± 0.023(↓ ↓)	0.070 ± 0.018(o ↓)	0.064 ± 0.023(↓)	$\mathbf{0.031 \pm 0.011}$
Movement	0.048 ± 0.012(o o)	0.058 ± 0.020(o ↓)	0.074 ± 0.028(↓ ↓)	0.089 ± 0.030(↓ ↓)	0.046 ± 0.009(o)	$\mathbf{0.043 \pm 0.013}$
Hillvally	0.033 ± 0.005(o ↓)	0.035 ± 0.007(↓ ↓)	0.037 ± 0.010(↓ ↓)	0.038 ± 0.009(↓ ↓)	0.031 ± 0.005(↓)	$\mathbf{0.026 \pm 0.005}$
Musk1	0.039 ± 0.004(↓ ↑)	0.025 ± 0.005(↓ ↑)	0.029 ± 0.007(↓ ↑)	0.029 ± 0.006(↓ ↑)	$\mathbf{0.022 \pm 0.004}$(↑)	0.043 ± 0.006
Multiple	0.140 ± 0.008(↓ ↓)	0.062 ± 0.020(↓ ↑)	0.057 ± 0.018(↓ ↑)	0.134 ± 0.014(↓ ↓)	$\mathbf{0.044 \pm 0.015}$(↑)	0.092 ± 0.013
Arrhythmia	0.114 ± 0.007(↓ ↓)	0.033 ± 0.009(o o)	0.051 ± 0.020(↓ ↓)	0.043 ± 0.012(o ↓)	0.038 ± 0.012(↓)	$\mathbf{0.030 \pm 0.014}$
SRBCT	0.226 ± 0.005(↓ ↓)	$\mathbf{0.023 \pm 0.016}$(↑ ↑)	0.141 ± 0.010(↓ ↓)	0.055 ± 0.020(↓ o)	0.039 ± 0.010(↑)	0.056 ± 0.008
Leukemia	0.214 ± 0.010(↓ ↓)	0.087 ± 0.036(o ↓)	0.172 ± 0.015(↓ ↓)	0.108 ± 0.034(o ↓)	0.087 ± 0.025(↓)	$\mathbf{0.070 \pm 0.007}$
DLBCL	0.177 ± 0.004(↓ ↓)	$\mathbf{0.028 \pm 0.018}$(↑ ↑)	0.135 ± 0.008(↓ ↓)	0.086 ± 0.017(↓ ↓)	0.060 ± 0.009(↑)	0.073 ± 0.005
Ranking	(2/6/6) (1/2/11) 3.93	(2/8/4) (4/1/9) 3.14	(0/1/13) (2/0/12) 4.82	(0/4/10) (1/1/12) 4.75	(4/1/9) 2.43	1.99

B. GF-NSGAII vs Others

Now we will focus on analyzing the effect of the grid-dominance. Fig. 3 shows that by using the proposed subset filtration mechanism and grid-dominance, the number of unique feature subsets in GF-NSGAII is slightly smaller than that of F-NSGAII during the evolutionary process. The reason is that when using the relaxed form of the Pareto-dominance, i.e., the grid-dominance, multiple feature subsets may sit within the same grid in the objective space. By using these solutions in crossover and mutation, some duplicated solutions may be produced. For the three large datasets (the SRBCT, Leukemia, DLBCL datasets), the number of unique feature subsets between F-NSGAII and GF-NSGAII is almost the same, but both methods have a significantly higher number of unique feature subsets than NSGAII.

In Tables II and III, the second sign in the brackets presents the significance test results, which compares each of the five algorithms with GF-NSGAII. Meanwhile, the second bracket in the last row of Tables II and III gives the sum of the significance test results. As can be seen from the last row of Tables II and III, GF-NSGAII shows the first performance among the six algorithms. More specifically, GF-NSGAII achieves significantly better HV results than GDE3 and NSGAII in all the used datasets. However, F-NSGAII achieves better HV and IGD results than the proposed GF-NSGAII algorithm on four datasets, namely Musk1, Multiple, SRBCT, and DLBCL.

As can be seen in Figs. 4 and 5, the obtained feature subsets from GF-NSGAII achieve a lower classification error rate than F-NSGAII on test sets, although the shapes of fronts between F-NSGAII and GF-NSGAII are similar on training sets. Let take the fronts on the SPECT dataset as an example. GF-NSGAII and F-NSGAII contain 27 feature subsets in their respective front on the training set of SPECT, and they show a similar front shape. Meanwhile, the test error rate of the obtained solutions from F-NSGAII varies in the range [0.222, 0.309], while that range for GF-NSGAII is mainly in [0.140, 0.224]. The main reason is that the grid-dominance can help an EMO based feature selection algorithm to obtain good feature subsets. These good feature subsets from GF-NSGAII include different features compared with the solutions obtained by F-NSGAII, while some feature subsets from F-NSGAII and GF-NSGAII can achieve similar or even the same training performance (i.e., the reason that GF-NSGAII and F-NSGAII show a similar front shape on the training set of SPECT) but may show different test performance.

In summary, GF-NSGAII won 97, draw 23 and lost 20 out of the 140 comparisons. The results of GF-NSGAII indicate that by applying the grid-dominance and the proposed subset filtration mechanism, GF-NSGAII can achieve better HV and IGD results. More importantly, GF-NSGAII can find effective and different feature subsets which show similar or even the same training performance as the benchmark algorithms, but

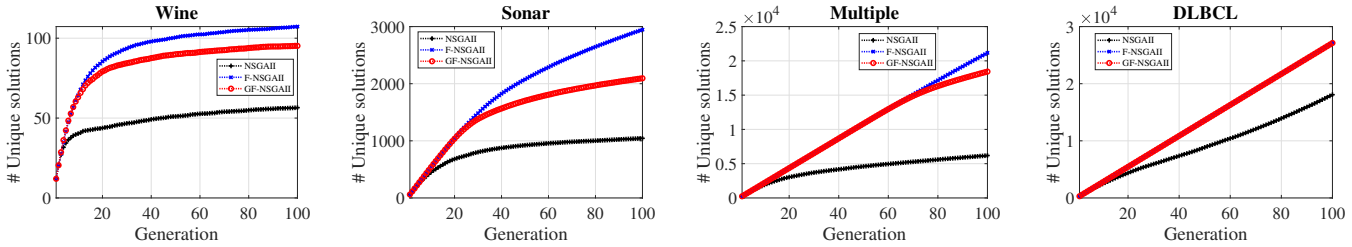

Fig. 3: The number of unique feature subsets during the evolutionary training process.

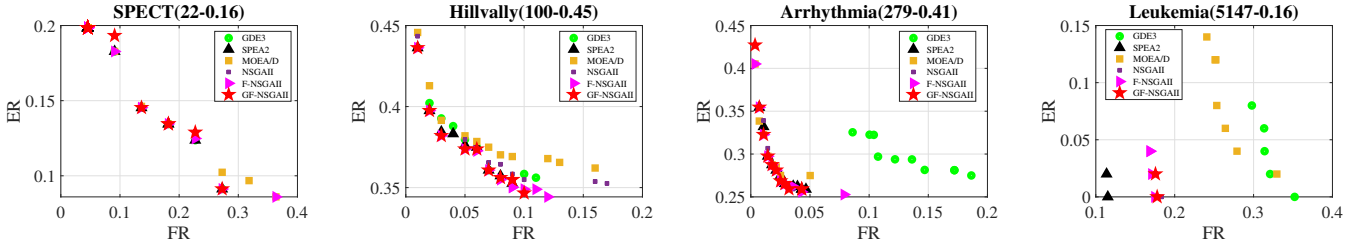

Fig. 4: The obtained PFs of the six algorithms on the training sets.

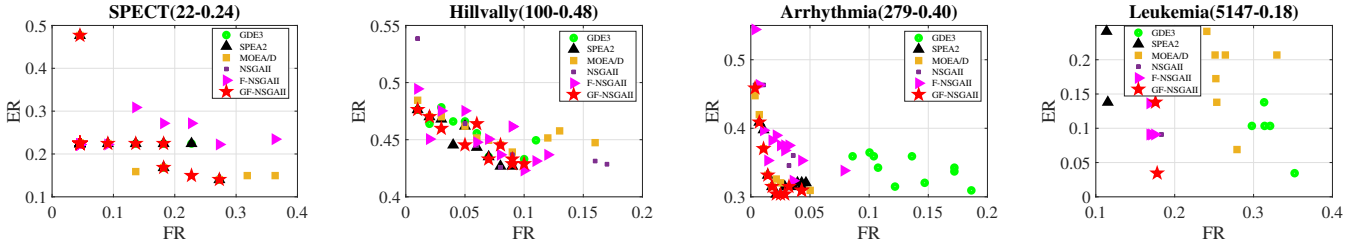

Fig. 5: The obtained PFs of the six algorithms on the test sets.

these feature subsets can show better test performance.

C. Computational Time

The average computational time (in minutes) of the six algorithms are shown in Fig. 6. As shown, the slowest algorithm among all the six compared methods is SPEA2, i.e., SPEA2 consumes the longest time on 11 out of the 14 datasets. For example, the average training time (177 minutes) of SPEA2 on the Arrhythmia dataset is more than five times that of other methods (less than 30 minutes). This may be due to the environmental selection strategy of SPEA2 calculating distances between solutions and using the information of the neighbours of the target solution.

On the small and medium datasets, the average running time of F-NSGAII and GF-NSGAII are almost the same as GDE3, MOEA/D, and NSGAII. On the three large datasets (the SRBCT, Leukemia, and DLBCL datasets), F-NSGAII and GF-NSGAII take a longer time than other four algorithms.

In comparison with F-NSGAII, GF-NSGAII is a little bit faster on almost all the used datasets apart from the WBCD and Multiple datasets. The most significant difference is on the DLBCL dataset, GF-NSGAII (148 minutes) saves nearly 30 minutes of the training time over F-NSGAII (120 minutes). However, the differences in the running/training time between the two algorithms on the other datasets are small.

In summary, F-NSGAII and GF-NSGAII require a slightly more running time than the other compared methods except for SPEA2.

VI. Conclusions and Future Work

This study aimed to design a grid-dominance based multi-objective feature selection algorithm to improve the classification performance. The goal has been successfully achieved by proposing a subset filtration mechanism and using the concept of grid-dominance. The proposed method was examined by incorporating it into the representative EMO approach (i.e., NSGAII) to form two different NSGAII based feature selection algorithms, and they were compared with the other four commonly used multi-objective algorithms on fourteen datasets. The results showed that by employing the proposed subset filtration mechanism, the population diversity was significantly increased. More importantly, the results showed that the grid-dominance can help the NSGAII based feature selection method to find more effective and different feature subsets, and achieved remarkably better HV and IGD results than GDE3, SPEA2, MOEA/D, NSGAII, and F-NSGAII.

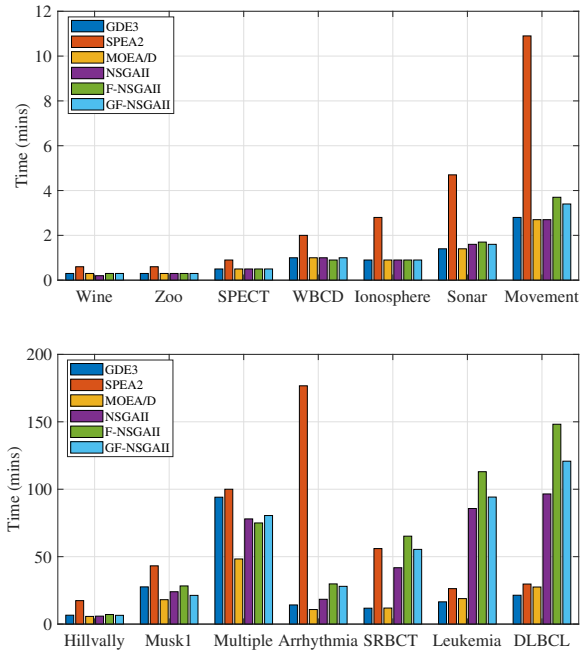

Fig. 6: The running time of the six algorithms on the training process.

In the future, we will further investigate new evolutionary mechanisms to avoid producing duplicated solutions to maintain the population diversity. Also, we will consider using a number of measures, e.g., information based measures, to design a local search strategy since although some feature subsets can achieve similar or the same objective values, there may still exist redundancy in those feature subsets.

ACKNOWLEDGEMENT

This work was supported in part by the Marsden Fund of New Zealand Government under Contracts VUW1509, VUW1615, VUW1913 and VUW1914, the Science for Technological Innovation Challenge fund under grant E3603/2903, the University Research Fund at Victoria University of Wellington grant number 223805/3986, MBIE Data Science SSIF Fund under the contract RTVU1914, and National Natural Science Foundation of China under Grant 61876169. This work of Peng Wang was supported by China Scholarship Council and Victoria University Scholarship.

REFERENCES

[1] R. Gilad-Bachrach, A. Navot, and N. Tishby, "Margin based feature selection-theory and algorithms," in *Proceedings of the Twenty-first International Conference on Machine Learning*, 2004, p. 43.

[2] B. Xue, M. Zhang, W. N. Browne, and X. Yao, "A survey on evolutionary computation approaches to feature selection," *IEEE Transactions on Evolutionary Computation*, vol. 20, no. 4, pp. 606–626, 2015.

[3] H. Liu and H. Motoda, *Feature extraction, construction and selection: a data mining perspective*. Springer Science & Business Media, 1998, vol. 453.

[4] B. Tran, M. Zhang, and B. Xue, "A pso based hybrid feature selection algorithm for high-dimensional classification," in *IEEE Congress on Evolutionary Computation*, 2016, pp. 3801–3808.

[5] B. H. Nguyen, B. Xue, P. Andreae, H. Ishibuchi, and M. Zhang, "Multiple reference points-based decomposition for multiobjective feature selection in classification: static and dynamic mechanisms," *IEEE Transactions on Evolutionary Computation*, vol. 24, no. 1, pp. 170–184, 2019.

[6] C. Yue, J. Liang, B. Qu, K. Yu, and H. Song, "Multimodal multiobjective optimization in feature selection," in *IEEE Congress on Evolutionary Computation*, 2019, pp. 302–309.

[7] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[8] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: Nsga-ii," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.

[9] E. Zitzler, M. Laumanns, and L. Thiele, "Spea2: improving the strength pareto evolutionary algorithm," *TIK-report*, vol. 103, 2001.

[10] Q. Zhang and H. Li, "Moea/d: A multiobjective evolutionary algorithm based on decomposition," *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 6, pp. 712–731, 2007.

[11] M. Li, S. Yang, and X. Liu, "Shift-based density estimation for pareto-based algorithms in many-objective optimization," *IEEE Transactions on Evolutionary Computation*, vol. 18, no. 3, pp. 348–365, 2013.

[12] C. Yue, J. Liang, B. Qu, Y. Han, Y. Zhu, and O. D. Crisalle, "A novel multiobjective optimization algorithm for sparse signal reconstruction," *Signal Processing*, vol. 167, p. 107292, 2020.

[13] H. Xu, B. Xue, and M. Zhang, "A duplication analysis based evolutionary algorithm for bi-objective feature selection," *IEEE Transactions on Evolutionary Computation*, DOI: 10.1109/TEVC.2020.3016049.

[14] S. Yang, M. Li, X. Liu, and J. Zheng, "A grid-based evolutionary algorithm for many-objective optimization," *IEEE Transactions on Evolutionary Computation*, vol. 17, no. 5, pp. 721–736, 2013.

[15] J. Cheng, G. G. Yen, and G. Zhang, "A grid-based adaptive multi-objective differential evolution algorithm," *Information Sciences*, vol. 367, pp. 890–908, 2016.

[16] L. Li, G. Li, and L. Chang, "A many-objective particle swarm optimization with grid dominance ranking and clustering," *Applied Soft Computing*, vol. 96, p. 106661, 2020.

[17] H. B. Nguyen, B. Xue, P. Andreae, and M. Zhang, "Particle swarm optimisation with genetic operators for feature selection," in *IEEE Congress on Evolutionary Computation*, 2017, pp. 286–293.

[18] K. Deb, S. Agrawal, A. Pratap, and T. Meyarivan, "A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: Nsga-ii," in *International Conference on Parallel Problem Solving from Nature*. Springer, 2000, pp. 849–858.

[19] P. Wang, B. Xue, J. Liang, and M. Zhang, "Improved crowding distance in multi-objective optimization for feature selection in classification," in *24th International Conference on Applications of Evolutionary Computation*, 2021, p. 489–505.

[20] N. Hallam, P. Blanchfield, and G. Kendall, "Handling diversity in evolutionary multiobjective optimization," in *IEEE Congress on Evolutionary Computation*, vol. 3, 2005, pp. 2233–2240.

[21] S. Kukkonen and J. Lampinen, "Gde3: the third evolution step of generalized differential evolution," in *IEEE Congress on Evolutionary Computation*, vol. 1, 2005, pp. 443–450.

[22] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[23] A. Benitez-Hidalgo, A. J. Nebro, J. Garcia-Nieto, I. Oregi, and J. Del Ser, "jmetalpy: A python framework for multi-objective optimization with metaheuristics," *Swarm and Evolutionary Computation*, vol. 51, p. 100598, 2019.

[24] B. Tran, B. Xue, and M. Zhang, "Variable-length particle swarm optimization for feature selection on high-dimensional classification," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 3, pp. 473–487, 2018.

[25] H. A. Shehu, A. Siddique, W. Browne, and H. Eisenbarth, "Lateralized approach for robustness against attacks in emotion categorization from images," in *24th International Conference on Applications of Evolutionary Computation*, 2021, pp. 469–485.

[26] L. While, P. Hingston, L. Barone, and S. Huband, "A faster algorithm for calculating hypervolume," *IEEE Transactions on Evolutionary Computation*, vol. 10, no. 1, pp. 29–38, 2006.

[27] H. Ishibuchi, H. Masuda, and Y. Nojima, "Sensitivity of performance evaluation results by inverted generational distance to reference points," in *IEEE Congress on Evolutionary Computation*, 2016, pp. 1107–1114.