An evolutionary computation-based approach for feature selection

Fateme Moslehi[1] · Abdorrahman Haeri[1] (ID)

Abstract

Feature selection plays an important role in the classification process to decrease the computational time, which can reduce the dimensionality of a dataset and improve the accuracy and efficiency of a machine learning task. Feature selection is a process that selects a subset of features based on the optimization criteria. Traditional statistical methods have been ineffective for two reasons, one being to increase the number of observations and the other to increase the number of features associated with an observation. Feature selection methods are a technique to reduce computational time, a better understanding of data, and improve the performance of machine learning and pattern recognition algorithms. The proper definition for solving the feature selection problem is to find a subset of minimum features so that it has the sufficient information for the purpose of problem and to increase the accuracy of the classification algorithm. Several techniques have been proposed to remove irrelevant and redundant features. In this paper, a novel feature selection algorithm that combines genetic algorithms (GA) and particle swarm optimization (PSO) for faster and better search capability is proposed. The hybrid algorithm makes use of the advantages of both PSO and GA methods. In order to evaluate the performance of these approaches, experiments were performed using seven real-world datasets. In this paper the gain ratio index is used to rank the features. The efficiency of the developed hybrid algorithm has been compared with the applicability of the basic algorithms. The results collected over real-world datasets represent the effectiveness of the developed algorithm. The algorithm was examined on seven data sets and the results demonstrate that the presented approach can achieve superior classification accuracy than the other methods.

Keywords Feature selection · Evolutionary approach · Genetic algorithm · Particle swarm optimization (PSO) · Gain ratio index

1 Introduction

Over recent years a set of high dimensional data and relatively low patterns have been produced along with the progress in science and technology. The goal of data mining process and knowledge discovery is to achieve intelligent decisions from a massive amount of data (Chen et al. 2011; Anagaw and Chang 2018). A large number of unrelated and inappropriate features reduces the learning speed and learning models. The growth of increases the complexity of building models is caused by the problem of is titled as the "curse of dimensionality" in data mining. This means

irrelevant and redundant features have a negative effect and decrease the accuracy of the classification process (Gu et al. 2018). One of the principal means to decrease the dimensions of the data set is feature selection which prepares an effective subset of features that include the greatest amount of information for making learning models. The essential goal of the feature selection process is the selection of a subset of outstanding and important features by eliminating the unrelated features and the lesser amount of information of the set. This reduction in dimensions helps to improve the learning process speed and it creates some simple and intelligible predicting models and it prevents over fitting (Guyon and Elisseeff 2003; Liu and Yu 2005; Unler et al. 2011; Moradi and Gholampour 2016).

Data classification is an important task in the field of data mining, which can be used to mine the model of important data (Yang et al. 2017). This technique predicts the label of each record according to the information on different features. In many issues, the data set classification consists of

✉ Abdorrahman Haeri
ahaeri@iust.ac.ir

Fateme Moslehi
moslehi_fateme@ind.iust.ac.ir

1 School of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran

a great number of features. However, all of these features are not useful in classification algorithm because the irrelevant and redundant features reduce the efficiency and precision of the classification algorithms. Therefore, in many cases, the high dimensions of the data are one of the serious obstacles in the way of classification (Guyon and Elisseeff 2003). In the feature selection process, many relevant features have been selected from among the whole primary features of the data set. The subset of the feature should have enough and qualified information to define the intended objective. Irrelevant features can decelerate the process of data mining without providing additional beneficial effect (Ghadimi et al. 2018). The feature selection advances the precision and performance of the classification algorithms by eliminating the absurd and repetitive features and speeds up the learning process (Dash and Liu 1997; Xue et al. 2012). Extracting important features has become a necessary phase for data mining to avoid the curse of dimensionality and reduce the amount of data for processing (Rao et al. 2019).

The feature selection techniques are generally separated into three classes: wrapper, embedded, and filter methods (Guyon and Elisseeff 2003). The wrapper method (Kohavi and John 1997) employs the machine learning algorithms as a part of the evaluation function of the features subset. The method efficiency and desirability of the selected subset are assessed usually by a classification algorithm in these models. However these methods may achieve an acceptable result, the complexity of instructing the classifier of the high dimensions data sets is a challenge. The embedded methods (Lal et al. 2006) mix the knowledge of the class structure of the functions using a machine learning technique. The computational cost in these methods is less than those of wrappers; however, their speed is lower than filter approaches and also the selected subset depends on the learning machine. In the filter methods is a perfect independency between the learning machine and data. Hence, they apply a criterion without using the learning algorithm to evaluate the selected subset. The filter approaches (Duch et al. 2003) are relatively stronger than overfitting methods; however, they may be unable to extract the best subset for the classification or regression (Vergara and Estévez 2014). Compared to filter approaches, wrappers are more precise because they consider the relationships between their features. Although they are computationally more expensive than filters, and their function is heavily dependent on the applied learning algorithm (Pashaei and Aydin 2017; Faris et al. 2018).

The relations and interactions among the features and extensive search space have hanged the feature selection into one of the most difficult processes of data mining and classification (Xue 2013; Lane et al. 2013). Two, three or more features may be interrelated. It is also possible that one feature does not have an effect on the target variable, but it may be effective along with another feature. In addition, a

feature may be appropriate to the target when it is alone, but it may be considered to be a redundant feature along with other features. The other challenge is the extensive search space. There are 2n possible states for n features and all these existing solutions are investigated in most issues. However, some searching solutions, such as sequential forward and backward feature selection (SFS, SBS) (Whitney 1971; Marill and Green 1963), have been developed to solve this issue, but it is possible to get stuck in the local minimum by the implementation of these methods and the computations may also prolong (Hancer et al. 2015).

The main purpose of this research is to propose a hybrid multi-objective feature selection algorithm to classification problems to attempt to select a small subset of features and obtain higher prediction results than using all features. The GA (Tao et al. 2019; Malakar et al. 2019), PSO (Xue et al. 2013; Moradi and Gholampour 2016) and ACO (Tabakhi and Moradi 2015) are commonly used the meta-heuristic algorithm in the field of feature selection. Investigators are focusing on GA because of its simplistic style whereas PSO and ACO are more accurate in analogous functions. PSO comes with some striking features in comparison to GA, including memory, which allows retaining the information of good solutions by the entire particles. In GA, on the other hand, earlier information of the problem is demolished upon changing the population. It cooperates constructively between particles for sharing their knowledge. Besides, the PSO algorithm has an easy implementation, less adaptable factors, and affordable computation in both speed and memory supplies. A universal search approach to find a better solution in the feature selection task is provided by PSO, though, it is infested with two deficiencies of premature convergence and weakness in calibrating near local optimum points. To overcome such weaknesses, a novel hybrid feature selection method based on PSO and GA called HGP-FS is proposed. The present study aims at hybridizing PSO and genetic algorithms to achieve their fantabulous features through their synthesis. To be further specific, PSO presents a critically advantageous local searching, which is not present in GA. In GA, however, a universal viewpoint is considered with functioning on the whole population from the initial starting point. Thus, PSO and GA can annul one another's disadvantages in the event of hybridization. In the HGP-FS, the fitness function is designed by using ANN algorithm. In the proposed approach the gain ratio index is applied for preprocessing and ranking the features and some features are omitted in this step. In general, the proposed method benefits from combining three powerful algorithms that perform well in selecting the appropriate subset of features in comparison to other similar works. The feature selection algorithm will be inspected and compared with the basic algorithms GA, PSO and ACO evolutionary algorithms on seven benchmark data sets with different

numbers of features, classes, and instances (details are shown in Sect. 4). Specifically, this paper will be investigated the following:

- Whether the use of ANN as the fitness function can succor GA-PSO to derive smaller feature subsets with higher classification precision,
- Whether combining the GA and PSO algorithms can achieve better classification performance and scalability than using all instances and a state-of-the-art sequential feature selection algorithm.
- Whether applying the method of the gain ratio will increase the performance of multi-objective feature selection algorithm and more effectively remove redundant or irrelevant features than the seven existing algorithms,

1.1 Related works

Nemati et al. (2009) have introduced an efficient feature selection approach for selecting useful features and elimination the irrelevant ones in biological data in a top-down way. The suggested approach integrates genetic algorithms (GA) and ant colony optimization (ACO) for using of benefits of both ACO and GA methods. The proposed approach had better performance in comparison with the basic algorithms of GA and ACO. Either ACO or GA is employed to generate the best solution. Thus, ACO can use the cross-over and mutation operations of GA. This can result in exploring ants around the best solution found in next iterations. The process iterates once again after phero-mone the pheromone is updated. Gheyas and Smith (2010) have introduced an efficient feature selection approach that combined genetic algorithms and simulated annealing. Their approach (SAGA) applies generalized regression neural networks (GRNN) classifiers until assessing each candidate feature subset solution. This approach was a solution for solving the problem of trapping in a local minimum of simulated annealing. Yang et al. (2011) have developed a feature subset selection technique by using genetic algorithm. The chromosome was split into several sections depending on the number of feature groups in their approach. Also, the value of fitness set up the probability of crossover and mutation. The present study proposes an improved genetic algorithm (IGA) by which an optimal feature subset has been chosen effectively and efficiently from a multi-character feature set (MCFS). The segmented chromosome management scheme for implementation of local management of chromosome is adopted in IGA. Xue et al. (2013) have proposed a wrapper algorithm for selecting useful features and elimination the inappropriate ones by using particle swarm optimization (PSO) algorithm. In the present study, two PSO-based multi-objective feature selection algorithms are

investigated. The idea of non-dominated sorting into PSO to solve feature selection problems is introduced by the first algorithm. The second algorithm employs the PSO based on the ideas of mutation, crowding and dominance. Bostani and Sheikhan (2017) have developed a hybrid feature selection algorithm by adopting the binary gravi-tational search algorithm (BGSA). The presented algo-rithm, known as MI-BGSA, performs the global search by BGSA as a wrapper-based feature selection method. Furthermore, the mutual information method aligns with the BGSA act as a filter-based phase, to reduce the feature subsets. Whale optimization algorithm (WOA) was applied to develop a novel wrapper feature selection method by Mafarja and Mirjalili (2018). In this study, two hybridiza-tion models are applied to design various feature selection methods using WOA. The novelty of these approaches are combining the SA algorithm and global search of WOA. In the first model, simulated annealing (SA) algorithm is embedded inside WOA algorithm, the aim of applying the SA algorithm is enhancing the best solution found in each iteration of WOA algorithm in the other approach. Mafarja et al. (2018) have proposed a hybrid algorithm based on Grasshopper optimization algorithm (GOA) and evolutionary population dynamics (EPD) methods. They introduced four different techniques by applying GOA and EPD. One of the upper three agents and one produced at random are selected in the first two methods to dislodge a solution from the poorest half of the population. In the third and fourth methods, roulette wheel selection (RWS) and tournament selection (TS) are applied for selecting the conducting agent from the first half in order to pro-vide an opportunity for low fitness solutions to reform the population. The performance of the proposed algorithm was better than the other similar methods in terms of the size of the selected subset, consumed CPU time and clas-sification accuracy. Hancer et al. (2018) have developed a novel hybrid feature selection algorithm with using binary artificial bee colony (ABC) algorithm. Two different repre-senting of ABC algorithm was applied; ABC with binary representation and ABC with continuous representation. The results achieved over real-world datasets show the efficiency of the suggested method. The algorithm was examined on 12 benchmark datasets and the results outline the performance superiority of the proposed algorithm to previous algorithms. Bidgoli et al. (2019) have developed a new feature selection algorithm for multi-label learning problems. They applied the binary version of generalized differential evolution (BGDE3) in the proposed method. Yao et al. (2017) proposed a novel algorithm with the aim of addressing the challenge of existed in directly embed-ding locally linear embedding (LLE) into the framework of the graph-preserving feature selection. Clearly, each feature is constructed the difference between structures of

the graphs and the original data is employed to evaluate the efficiency of each feature. In Labani et al. (2017), a new multi-objective algorithm, known as multi-objective mutual information (MOMI) has been proposed by using mutual information. The strategy of this algorithm is minimizing the redundancy of features and maximize the relevancy with the target class. A feature selection algorithm for tumor segmentation and classification from MR images has proposed and developed by Sharif et al. (2018). A Gaussian filter addresses the preprocessing problem in the first step by implementing. Next, the GA is employed to choose the best features from a fused vector which are later on fed into LSVM for classification. Two powerful algorithms were used by Jayaraman and Sultana (2019) to develop a novel wrapper feature selection method they integrated the artificial gravitational cuckoo search algorithm and particle bee optimized associative memory neural network approaches. A new hybrid (wrapper-filter) approach by using the evolutionary algorithms has introduced by Moslehi and Haeri (2019). The main objective of this method is to reduce the complication of calculating and the search time to attain the important and sufficient subset of features. The reviewed works are summarized in Table 1.

The rest of the paper is structured as follows. Section 2 describes a brief overview of feature selection and evolutionary algorithms. Section 3 introduces the proposed approach. Section 4 presents the experimental results and discussions. Section 5 concludes the study and provides an insight into the future trends.

Table 1 The summary of related works

| References	Applied algorithm	Wrapper/ filter/ hybrid
Nemati et al. (2009)	ACO and GA	Wrapper
Gheyas and Smith (2010)	GA and SA	Wrapper
Yang et al. (2011)	IGA	Wrapper
Xue et al. (2013)	PSO	Wrapper
Bostani and Sheikhan (2017)	BGSA	Hybrid
Mafarja and Mirjalili (2018)	WOA and SA	Wrapper
Mafarja et al. (2018)	GOA and EPD	Wrapper
Hancer et al. (2018)	ABC	Hybrid
Bidgoli et al. (2019)	BGDE3	Wrapper
Labani et al. (2017)	Mutual information	Filter
Yao et al. (2017)	Graph Method	Filter
Sharif et al. (2018)	Gaussian filter and GA	Wrapper
Jayaraman and Sultana (2019)	Cuckoo search and particle bee optimized associative memory	Wrapper
Moslehi and Haeri (2019)	GA and PSO	Wrapper

2 Preliminaries

The definition of feature selection is the procedure that picks out a small subset of the relevant features from a set of main features that can decrease the data dimensions and enhance the performance of a learning technique. The progress in the techniques of data collection has increased the number of data set features and this has increased the feature selection importance in data mining science and machine learning (Lane et al. 2014). Different methods have been employed by the researchers to proposed solutions for the challenge of feature selection, and the evolutionary algorithms have attracted much attention from many researchers because of their search capability (Cervante et al. 2012).

2.1 Feature selection

Feature selection aims to select a set of features from the main data set that can explain the input data effectively while decreasing the effect of noise or inappropriate features and achieves high classification accuracy (Chandrashekar and Sahin 2014; Guyon and Elisseeff 2003). Figure 1 describes the general process of feature selection that includes four stages (Dash and Liu 1997)

1. Subset detection: The detection stage for the production of candidate subsets (Langley 1994). This stage is the search process stage that can start without any feature, using all variables or with a subset of random variables. In this stage, the search methods include the common methods and EC techniques to produce the feature subsets.
2. Subset assessment: an assessment criterion employed to compute the degree of desirability of the produced feature subsets.
3. Stopping condition: the implementation of the algorithm continues to the realization of the stopping condition, and it terminates upon the realization of the stopping situation. The disconnecting condition can be defined
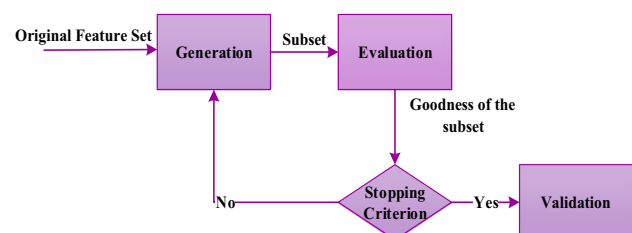
**Fig. 1** Feature selection process (Langley 1994)

based on an evaluation function, for instance, the counter of algorithm repetitions or the number of selected variables.

4. Validity: this stage focuses on the accreditation of the selected subset. This accreditation is performed through the experiment and analysis of the results of the suggested algorithm with former methods using the synthetic data sets or real-world datasets.

A subset of variables can be found by several search algorithms by maximizing the objective function that is the classification performance. The tree structure is the strategy of the Branch and Bound techniques for assessing the desirability of the subset of features. However, the search would have an exponential growth for the greater number of features. In recent years, meta-heuristic algorithms have gained considerable ground owing to their proper functioning to solve the feature selection issue. A huge body of searching approaches can have intensive computations for greater datasets. Hence simplified algorithms such as sequential search or evolutionary algorithms including Genetic algorithm (GA) or Particle swarm optimization (PSO) that can achieve local optimum solutions are chosen which can generate appropriate solutions and are computationally feasible (Chandrashekar and Sahin 2014). In this study GA and PSO are employed to develop the new feature selection method and these algorithms are introduced in the following.

2.2 Particle swarm optimization

Kennedy and Eberhart (1995) proposed PSO as a further instance of a population-based optimization algorithm, which was later undergoing developments in a variety of variants for testing problems and applications. The key metaphor utilized in particle swarm optimization is that a group of particles takes advantage of their personal and social experience for exploring a decision space and detecting high-performance solutions (Kıran et al. 2012). In this algorithm, like evolutionary computation techniques, the genetic algorithm starts with a random population that forms potential solutions and then the population is updated. In this algorithm, each response is called a particle, and features of this algorithm can be shared information by particles. The particles when repeating the algorithm, based on their current position information, are positioned in the best status (with the least prediction error) and the best position of the entire particle changed and positioned. Figure 2 shows the way of displacement of particle A. Assuming that a particle is in position A, such that its velocity vector is in line with vector B, three limit states can be considered for the subsequent movement of the particle. In the first case, the particle retains its previous direction of motion and moves in the same direction. It will then reach point B. In the latter case, the particle moves to the best
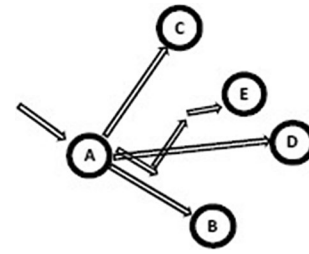
Fig. 2 How to change the location of the particles in the PSO algorithm

position experienced to reach point C. The third is moving in the direction of the best particle location and positioning D. The general state of motion in the algorithm POS and the new position of each particle can be calculated based on Eqs. (1) and (2) (Kıran et al. 2012).

$$V_i(t+1) = w \times V_i(t) + c_1 \times r_1 \times (Pbest_i(t) - X_i(t)) + c_2 \times r_2 \times (Gbest_i(t) - X_i(t)) \tag{1}$$

$$X_i(t+1) = X_i(t) + V_i(t+1) \tag{2}$$

Each particle has some possibilities including the best value it has ever reached ($Pbest$) and the X_i position. The information is obtained from the comparison of the attempts of each particle to find the best solution. Also, each particle represents the best solution ever obtained (in the group) from $Pbests$ value ($Gbest$). Each particle attempts to substitute its situation using the following information. The present position X, velocity V_i and the distance between the present position, $Pbest$, distance between the present position and $Gbest$. Velocity and position of particles are updated by the following Eqs. (1) and (2). Where $V_i(t)$, is the particle's velocity at time t, w is the inertia weight, c_1 and c_2 are weight factors, r_1 and r_2 are the random numbers between 0 and 1, $X_i(t)$ is the particle's position at time t, $Pbest_i$ is the ith dimension of each particle and $Gbest_i$ and $Gbest$ in the group. The PSO algorithm is summarized in the following steps:

1. Evaluate the fitness of each particle,
2. Update local best position,
3. Update global best position,
4. Update velocity of each particle,
5. Update position of each particle.

These steps are repeated until some stopping condition is met.

2.3 Genetic algorithm

Evolutionary computation was first presented in 1960 by Dr. Richenberg, whose research was on evolutionary strategy. His theory was later studied by many researchers

until the genetic algorithm was developed by John Holland at the University of Michigan in 1975 (Zeng et al. 2011). This algorithm starts with an acceptable response "k" known as the initial population and then, using the fitness function, calculates the utility of each response and then, from the available responses, two solutions are selected with desirable utility as a parent, and then new children are formed by mixing parent responses, and by adding them to the previous population, a new population is obtained (Zeng et al. 2011).

In general, genetic algorithms consist of the following components:

1. *Chromosome* In genetic algorithms, each chromosome represents a search space and a possible solution to the problem. Chromosomes (solutions) are made of a number of gene sequences. Binary coding (bit strings) is usually used to represent chromosomes
2. *Population* A series of chromosomes forms a population. With the influence of genetic functions on each population, a new set with the same number of chromosomes is formed.
3. *Fitness operators* In order to solve any problem using genetic algorithms, the fitness function must first be created for that problem. For each chromosome, this function returns a non-negative number indicating the suitability or individual ability of that chromosome.
4. *Genetic operators* Genetic algorithms use genetic operators during the reproductive phase. By influencing these operators on the population, the next generation of that population is produced. Selection, crossover, and mutation operators are most commonly used in genetic algorithms.

Genetic operators can be described as follows:

Selection operator This function selects a number of chromosomes for reproduction among the chromosomes in a generation. More graceful chromosomes are more likely to be selected for reproduction.

Crossover operator The crossover operator selects on a chromosome pair of generative generation and produces a new chromosome pair. There are several mixing functions, including single-point and two-point mixing. In single-point alignment, a random position is considered between two genes.

Mutation operator After the crossover is completed, the mutation operator is affected on the chromosomes. The purpose of mutation in GAs is preserving and introducing diversity. This operator selects a gene from a chromosome randomly and then changes the gene from its original state.

3 Proposed method

The data holding N records and F attributes are first loaded in this process. This procedure aims to determine which of these D attributes can be chosen as a suitable characteristic, so that the problem dimensions can be lowered in the first facet providing the original problem, viz. classification execution, is not hampered or disturbed, with increases in the speed and precision of classification. The general Schema of the feature selection problems is shown in Fig. 3.

It should, therefore, be ascertained that the classification precision can be maximized by selecting which of these D attributes. Genetic problem will generally have D dimensions when D features exist. To facilitate further, the above D dimensions can be taken as a binary collection holding D numbers in which zero means no selection of a feature and one means selection of a feature. Thus, it is speculated that population members for genetic algorithm and particle swarm optimization algorithm are as D dimension arrays whose values will be binary.

In meta-heuristic approaches, the key subject is the fitness function or the cost function. In this operation, this function is identified as the cost function and its input will be a N*M matrix, of which the output, viz. the classification error M value, is similar to a number of problem features (D) for which the classification cost is measurable. Notably, that $0 < M \leq D$. The value of M can never equal zero since a minimum of one feature needs to be chosen for the classification.

Initially, a preliminary analysis is performed on all the features of the dataset. The goal is to identify features that are more probably to be present in the selected subset of features. To do this the gain ratio technique is employed. This index is used in the C4.5 decision tree algorithm to select the features that make the most difference between the classes created at each step as the split feature for the decision tree branches. The gain ratio is a modification of the information gain that reduces its bias on high-branch attributes. The gain ratio takes number and size of branches into account when choosing an attribute. It corrects the
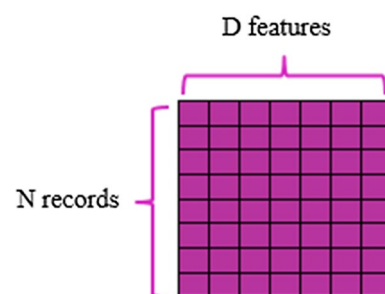

Fig. 3 The general schema of the problem

information gain by taking the *intrinsic information* of a split into account. The formula of this index is explained in the following Eqs. (3)–(6).

$$Gainratio = \frac{IG}{H(X)} \tag{3}$$

Where,

$$IG(X|Y) = H(X) - H(X|Y) \tag{4}$$

$$H(X) = -\sum_{i=1}^{k} P(x_i) \log_2 P(x_i) \tag{5}$$

Is the entropy of the variable X, and

$$H(X|Y) = -\sum_i P(y_j) \sum_i P(x_i|y_j) \log_2 \left(P(x_i|y_j) \right) \tag{6}$$

Is the entropy of X after observing another variable Y.

Features that have higher the gain ratio measure value means that they have more information value in the dataset, and thus are more likely to be present in the selected final subset. In other words, these are more important and more relevant features. Therefore, in this study, this criterion was used to pre-process the features at an early stage. This means that the gain ratio index is calculated for all attributes, and 15% of the features with the lowest gain ratio index are eliminated and will not be present during the following algorithm steps and the next steps will continue with only 85% features. This will remove the features that are less probably to be present in the selected subsets in the early stages. Therefore, the algorithm implementation speed as well as the prediction accuracy with the selected subset, will increase.

Even so, the considered cost function is under the control of an Artificial Neural Network (ANN) meaning that the designated data are delivered to a neural network that initiates classifying the data depending on the equivalent labels. Ultimately, the ANN returns an output as the classification error that is regarded as the level of cost (suitability value) for each member of the population. The problem seeks a member of the population that levies a minimum level of cost on the problem. As mentioned earlier, the proposed method is a multi-criteria method and several criteria are employed to choose the selected generation. The considered fitness function is described in Eqs. (3) and (4). Where *rf* is ratio of selected features, *E* is Average of *EE*, beta equals to 0.5, *wTrain* equals to 0.8 and *wTest* equals to 0.2. There are two key aims in the introduced procedures, viz. raising the prediction precision and lowering the numbers of selected features. Accordingly, these two essential measures (the prediction precision and the size of the selected subset) are regarded as objectives and surveyed

in the current research to assess the performance of this algorithm. Two points including numbers of features and prediction accuracy are important and effective in analysis and making decision. The high number of features will make the decision process difficult and the low prediction accuracy of the decision reduces the validity of the decisions made. The prediction accuracy has a direct effect in quality and strength of a decision. Therefore, both measures are important factors in the efficiency of decision tree and predictive methods. These two goals are achievable by an appropriate and proper feature selection algorithm in the preprocessing phase. As it is obvious in Eqs. (7) and (8) these two important measures are considered in the fitness function of the multi-objective proposed algorithm.

$$Fitness = E \times (1 + beta \times rf) \tag{7}$$

Where,

$$EE = wTrain \times Error\ of\ TrainData + wTest \times Error\ of\ TestData \tag{8}$$

As mentioned before, the population members need to have D dimensions. To this end, two populations with particular numbers and D dimensions are first created at random, one of which is for genetic algorithm and the other belongs to PSO algorithm. After generating the initial population and computing every member's cost, the technique begins running in three parallel stages and the procedure lasts by the time it reaches the end of the determined frequency number.

The above three steps operate in two sections of PSO and genetic algorithms. There are two parallel phases in the latter, and only one phase exists in the former.

The 1st phase (population generation through crossover): in the established method, a particular number of the population is initially chosen by roulette wheel algorithm. Next, crossover operations are run for every two members of the population (the two members are preferred to be named parents herein). Crossover is performed in three situations of single-point, two-point, and uniform, and one of these three conditions is selected at random in crossover procedures. In the 1st phase, therefore, crossover operations generate some offspring.

In the 2nd phase, mutation operations generate some distinct offspring. One member of the population is regarded in mutation operations, followed by doing mutation operation for the initially values with a probability. One point worthy of note is that only a specific number of the values may alter concerning the measured probability value. For instance, for a member with a value of 60 and a mutation probability of 0.2, only two values experience mutation operation and those values become reversed. Thus, some other offspring are produced by mutation operations in the 2nd phase.

Algorithm 1. The pseudo code of the HGP_FS algorithm.

1) Calculate the gain ratio index of each feature
2) Select the 85% of features with highest gain ratio value and Implement the following step on selected features:
3) Choose an initial random population of individuals
4) Evaluate the fitness of the individuals
5) Randomly initialize particles swarm
6) Evaluate Fitness of particles
7) Repeat
8) Select the best individuals to be used by the genetic operators
9) Generate new individuals using crossover and mutation
10) Evaluate the fitness of the new individuals
11) Update the velocity and position of particles
12) Update local best position and global best position of particles
13) Evaluate Fitness of particles
14) Integrate the values of population of steps 7 and 10
15) The integrated population are sorted according to their fitness value
16) Select a number of the best part of the integrated population
17) If the loop counter does not reach the specified limit, go to step 6
18) Extract the best solutions in the final integrated population
19) Regarding the final answer, get the specified features of the data
20) Use the selected features for training the classifier

The particle swarm optimization algorithm creates other offspring in the 3rd phase includes two bases: first, the best search of every member of the population and second, the best search present in the entire population. The population has moved towards getting the best solution during the time.

This algorithm is implemented together with GA, and in each movement, the population has moved towards creating new responses that are superior to the earlier ones. Following the motion of the population in a given route, the cost of every member of the population is computed by the cost function, followed by updating *Gbest* and *Pbest* of every member concerning the new costs. *Gbest* refers to the best response obtained by the total population hitherto, and *Pbest* denotes to the best response attained by every member of the population up to now. Thereafter, the responses are introduced to particle swarm optimization algorithm for the nest phase. After moving of the three phases, there will be eventually four populations: (a) the initial genetic population. (b) The population ensued from crossover. (c) The population stemmed from mutation. (d) The population arose from PSO movement.

At completion of each movement (cohort), the above four populations are incorporated and a specific number of them

is chosen for the next generation of genetic algorithm. This choice depends upon the cost and fittingness of the population. Consider that PSO population does not change.

The algorithm terminates after running of a specific number of the algorithm frequencies. The best answer is achieved in the final resultant population, and is announced as the best ultimate answer, which dictates selecting the features for the purpose of accuracy classification. The flowchart of the proposed algorithm is presented in Fig. 3 and all the phases of the algorithm are presented as bellow pseudo code in Algorithm 1.
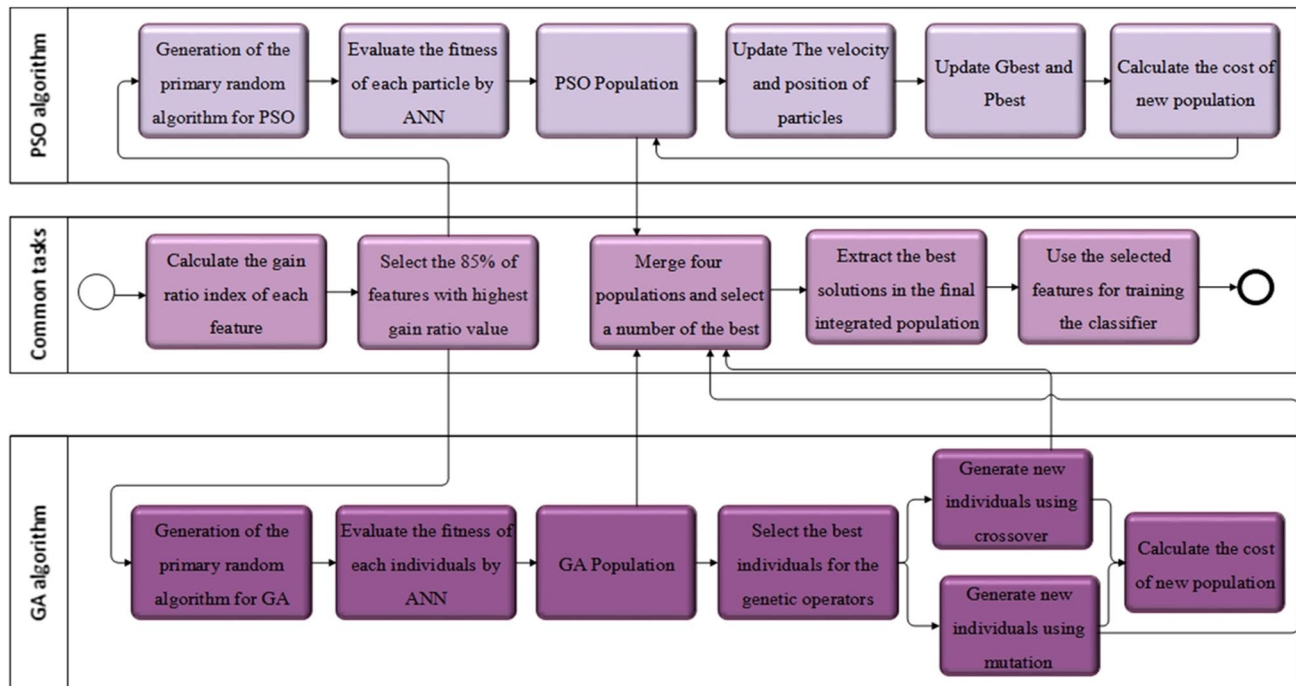
4 Experimental results

The experimental results are reviewed and survived in this section (Fig. 4).

4.1 Datasets

Several experiments were conducted to demonstrate the effectiveness of the proposed algorithm. In order to achieve this goal, seven benchmark datasets from the UCI machine learning repository are applied that their descriptions are

Fig. 4 The flowchart of proposed algorithm

Table 2 Datasets

Dataset	Number of features	Number of classes	Number of examples
Wine	13	3	178
German	24	2	1000
WBCD	30	2	569
Ionosphere	34	2	351
Lung	56	3	32
Hill-Valley	100	2	606
Musk 1	166	2	476

shown in Table 2. The datasets include a different number of variables, classes and examples that preparing a general and wide survey of the suggested and employed approaches. The parameters of the proposed algorithm are recorded in Table 3. All of the wrapper approaches to evaluate the classification efficiency use a classification algorithm in the evolutionary training process. Some classification techniques, like NB, DT, and SVM can be applied for this purpose. As a classifier, one of the popular algorithms k nearest neighbors

(KNN) was applied to measure the classification efficiency. In these experiments k is adjusted 5 (5NN).

4.2 Results

For more precise analysis, the proposed algorithm has been executed on every data sets 10 times, and the average results are recorded in Tables 4 and 5. In the experiments, all of the instances in each data set are randomly divided into two sets: 70% as the training set and 30% as the test set. The classification performance of a selected feature subset is assessed by tenfold cross-validation on the test set. Experimental results of six approaches on seven datasets are reported in Table 2. "Ave-Size" shows the mean size of the feature subsets selected by proposed algorithm in 10 implementations. In experimental results the mean test accuracy of the feature subsets selected by each algorithm is represented by "Ave-Acc", the best test precision is showed by "Best-Acc" and "Std-Acc" is used to indicate the standard deviation of the 10 test accuracies obtained by several runs. The first column in Table 4 is the results of the 5-NN with all variables and

Table 3 The parameter settings

Parameter	Population size	No. of generations	Crossover probability	Mutation probability	C1 and C2	W
Average	60	350	0.8	0.3	2	1

Table 4 Experimental results

Dataset	5-nn All	GA	BPSO	MRABC	BPSO-Er	BPSO-ErNo	BPSO-2Stage	ErFS	2SFS	HGP-FS
Wine										
Ave-size	13	6.1	5.8	5.76	8.32	8.1	*5.1*	8	8	6.44
Ave-acc	76.54	96.05	95.62	95.55	95.96	96.23	96.94	95.96	95.96	*96.94*
Best-acc					97.53	98.77	100	100	100	*100*
Std-acc		5E−1	9E−2	12E−2	1.87E−2	1.58E−2	2.51E−2	1.83E−2	1.83E−2	1.19E+0
German										
Ave-size	24	10.70	11.43	10.40	12.9	9.48	8.62	13.48	11.92	*5.08*
Ave-acc	68	75.07	75.2	75.55	68.73	68.83	68.93	69.41	69.15	*78.17*
Best-acc					72	71	73.67	72	72	*82.71*
Std-acc		7E−1	4.9E−1	6.1E−1	1.3E−2	1.35E−2	1.65E−2	1.33E−2	1.18E−2	4.9E+00
WBCD										
Ave-size	30	14.96	13.36	12.23	14.92	7.65	6.68	13.42	**5**	10.09
Ave-acc	92.98	95.28	96.01	95.47	92.98	92.98	92.98	93.39	93.54	*96.48*
Best-acc					92.98	92.98	92.98	94.74	94.74	*97.72*
Std-acc		1.46E+0	4.3+E1	3.9+E1	3.33E−1	3.33E−1	3.33E−1	55.8E−4	75.3E−4	8.4E−02
Ionosphere										
Ave-size	34	10.93	10.1	10.13	10.38	**8.55**	8.9	12.58	12.05	14.85
Ave-acc	83.81	91.28	91.11	91.24	89.05	89.12	89.52	88.4	88.14	*91.37*
Best-acc					93.33	*94.29*	93.33	93.33	91.43	92.5
Std-acc		1.75E+0	8E+1	1.05E+1	1.84E−2	1.84E−2	1.59E−2	2.14E−2	1.89E−2	8.53E−01
Lung										
Ave-size	56	28.03	27.33	26.50	26.92	23.5	*22.22*	27.35	27.38	*22.2*
Ave-acc	70	74.9	76.18	76.65	72.5	73	73.25	72	72.25	*86.18*
Best-acc					80	90	80	80	90	*95.83*
Std-acc		1.26E+01	3.95	5.5	5.36E−2	5.57E−2	6.08E−2	6E−2	6.89E−2	5.97E+00
Hill-Valley										
Ave-size	100	45.73	46.16	44.73	47.55	37.75	37.1	47.32	47.05	*36.25*
Ave-acc	56.59	59.17	56.09	58.21	57.56	57.72	57.61	57.54	57.57	*78.53*
Best-acc					60.71	60.16	60.44	61.81	61.81	*78.96*
Std-acc		2.05E+0	0.96	1.31	1.48E−2	1.36E−2	1.19E−2	1.52E−2	1.55E−2	8.5E−01
Musk 1										
Ave-size	166	81.83	81.2	82.26	83.6	**79.4**	80.72	86.48	85.58	91
Ave-acc	83.92	*92.58*	90.32	91.02	85.65	85.54	85.7	84.58	84.58	90.06
Best-acc					89.51	90.91	89.51	88.81	88.81	*95.73*
Std-acc		2.36E+0	0.55	0.73	2.1E−2	2.21E−2	2.05E−2	2.04E−2	1.99E−2	3.24E+00

no feature selection. The best result in each row is shown in bold format in Tables 4 and 5.

The proposed method (HGP-FS) is compared to state-of-the-art feature selection algorithms including BPSO-Er, BPSO-ErNo and BPSO-2Stage (Xue et al. 2012) ErFS and 2SFS (Xue et al. 2013) and GA, BPSO, MRABC (are derived from Hancer et al. (2015)). The K-NN method is one of the major and powerful classification algorithms and basis of this method is on a similarity measure. To assess the efficiency of the presented feature selection approach, the five nearest neighbors (5-NN) is applied.

The results of the implementation of the algorithm were compared with previous similar works which showed an improvement in the results from a couple of criteria's point of view including the average and standard deviations of precision rate and size of subset selected feature that are shown in Table 4. The efficiency and desirability of the examined algorithms are deduced from the accuracy of the 5-NN classifier accuracy. The suggested method has been evaluated with other approaches and as the results are shown in Table 4, the algorithm has higher efficacy and reliability compared to the similar algorithms and it can be stated that this procedure was successful in removing non-relevant features and decreased data dimensions with increased accuracy of classification at the same time. Although the proposed method was not successful in the selection of fewer subsets,

**Table 5** The obtained accuracy rates of the deterministic approaches with the proposed approach

Dataset	LFS	LFFS	GSBS	HGP-FS
Wine				
Ave-size	7	8	10	**6.44**
Ave-acc	74.07	74.07	75.93	**96.30**
Germanc				
Ave-size	**5**	**5**	20	*5.08*
Ave-acc	68.33	68.33	69.33	*78.17*
WBCD				
Ave-size	12	11	29	*10.09*
Ave-acc	92.35	91.18	92.94	*96.48*
Ionosphere				
Ave-size	*6*	*6*	29	14.85
Ave-acc	90.48	90.48	89.52	*91.37*
Lung				
Ave-size	6	*5*	36	36.25
Ave-acc	66.67	66.67	66.67	*78.53*
Hill-Valley				
Ave-size	9	8	95	*36.25*
Ave-acc	55.49	53.85	54.40	*78.53*
Musk 1				
Ave-size	*12*	*12*	124	91
Ave-acc	80.71	79.29	82.86	*90.06*

it showed a progression regarding the increase of accuracy compared to the previous algorithms In fact; this method decreased the characters of dataset dimensions compared to the other procedures. As the proposed method showed higher efficacy with both datasets with high dimensions like Hill-Valley and datasets with low dimensions like German, the reliability and stability of this procedure are emphasized. As shown in Table 4, the accuracy of the developed approach in the data sets Hill-Valley, WBCD, and German is
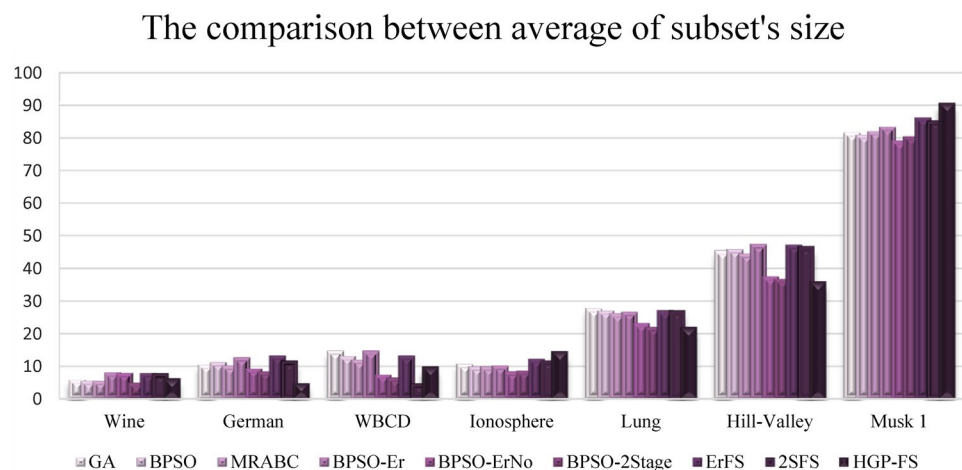
higher than other methods and the superiority and efficiency of the presented approach are evident compared to the other method. The obtained results demonstrated that proposed algorithm has better performance in terms of the size of the selected subset this means that HGP-FS method can select the subset of features with smaller size. Therefore, the algorithm has been successful in extracting the relevant subset feature more efficiently than other methods. As a result, it is obvious that this approach has a better efficiency in comparison with the other similar feature selection approaches. Although the accuracy rate results derived from the HGP-FS algorithm in the in Musk 1 data set was very close to other methods and had a minor difference. The conducted comparisons are presented in Figs. 5 and 6.

Moreover, the proposed method is compared the deterministic approaches including the linear forward selection (LFS) (Gutlein et al. 2009), linear floating forward selection (LFFS) (Gutlein et al. 2009), and greedy stepwise based selection (GSBS) (Caruana and Freitag 1994). The results of LFS, LFFS, and GSBS are derived from Hancer et al. (2015). The comparison between proposed approach and the deterministic approaches are shown in Table 5.

As it is obvious in Table 5 for all datasets the current approach has better performance in comparison to the rest of techniques in the terms of accuracy prediction. LFFS and LFS methods have chosen those datasets with fewer features, though, the categorization obtained from those datasets was found to be less accurate. GSBS algorithm has an aggressive function with the entire primary datasets in the procedure of selecting feature. Additionally, a subset with greater numbers of features is selected in this procedure and no enhancement was observed in an accurate categorization. Accordingly, the introduced method demonstrated greater effectiveness in comparison to deterministic approaches and

Fig. 5 The comparison between averages of accuracy

The comparison between average of accuracy

GA ▪ BPSO ▪ MRABC ▪ BPSO-Er ▪ BPSO-ErNo ▪ BPSO-2Stage ▪ ErFS ▪ 2SFS ▪ HGP-FS
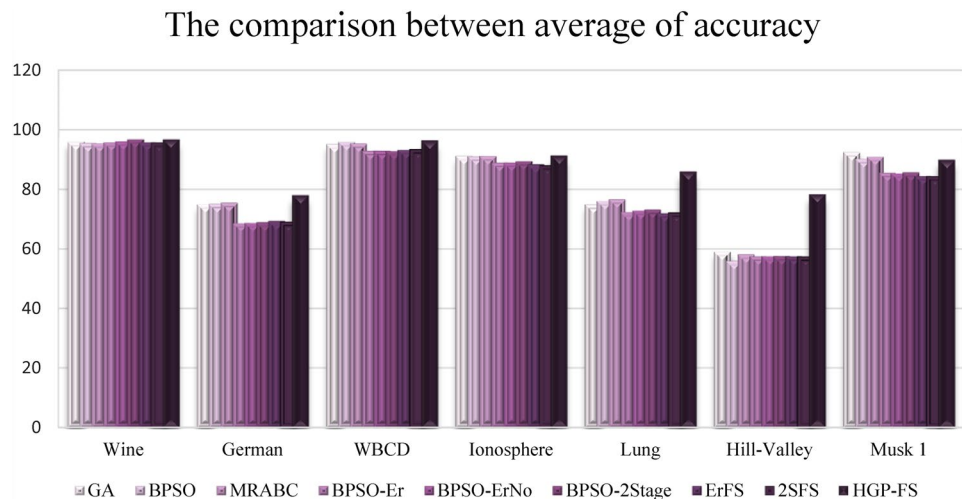
Fig. 6 The comparison between averages of subset's size

The comparison between average of subset's size

selected a subset that led to achieving an accurate categorization with better quality and a considerable enhancement in this field.

5 Conclusions

Feature selection is a powerful technique to reduce dimensionality and is an important step in the success of the machine learning process. Over the past few decades, the datasets have grown exponentially in both sample numbers and feature numbers, making it difficult to handle feature selection. Generally, the aim of integrating GA and PSO is to tradeoff between exploration and exploitation resulted in an optimum solution. The proposed procedure titled HGP-FS applied the algorithm in objective function. The present research has sought to create an algorithm without the previous faults by combining these algorithms and using their strengths that is the high convergence speed of particle swarm optimization algorithm and exploration ability of genetic algorithm. The obtained results show that the integration of GA and PSO algorithm effectively improved the global search ability of the feature selection process, and the proposed algorithm achieved the best classification performance in both training and test sets in almost all cases. To assess the efficiency of the suggested feature selection algorithm the obtained result of the experiments were compared to those of the state-of-the-art wrapper based feature selection methods like GA, BPSO-Er, BPSO-ErNo and BPSO-2Stage and the deterministic approaches. The results found from the comparison of the proposed algorithm and similar algorithms showed better effectiveness of the proposed algorithm to eliminate features and increase the accuracy of classification in both big and small datasets accounting for the stability of the proposed method. Further, the proposed algorithm can remove redundant features effectively while obtaining the highest classification results.

References

Anagaw A, Chang YL (2018) A new complement naïve Bayesian approach for biomedical data classification. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/s12652-018-1160-1

Bidgoli AA, Rahnamayan S, Ebrahimpour-Komleh H (2019) Opposition-based multi-objective binary differential evolution for multi-label feature selection. In: International conference on evolutionary multi-criterion optimization. Springer, Cham, pp 553–564. https://doi.org/10.1007/978-3-030-12598-1_44

Bostani H, Sheikhan M (2017) Hybrid of binary gravitational search algorithm and mutual information for feature selection in intrusion detection systems. Soft Comput 21(9):2307–2324

Caruana R, Freitag D (1994) Greedy attribute selection. In: Machine learning proceedings 1994. Morgan Kaufmann, pp 28–36

Cervante L, Xue B, Zhang M, Shang L (2012) Binary particle swarm optimisation for feature selection: a filter based approach. In: 2012 IEEE congress on evolutionary computation (CEC), pp 1–8. https://doi.org/10.1109/CEC.2012.6256452

Chandrashekar G, Sahin F (2014) A survey on feature selection methods. Comput Electr Eng 40(1):16–28

Chen Y, Miao D, Wang R, Wu K (2011) A rough set approach to feature based on power set tree. Knowl Based Syst. https://doi.org/10.1016/j.knosys.2010.09.004

Dash M, Liu H (1997) Feature selection for classification. Intell Data Anal 1(3):131–156

Duch W, Winiarski T, Biesiada J, Kachel A (2003) Feature selection and ranking filters. In: International conference on artificial neural networks (ICANN) and international conference on neural information processing (ICONIP), vol 251, pp 254–262

Faris H, Mafarja MM, Heidari AA, Aljarah I, Ala'M AZ, Mirjalili S, Fujita H (2018) An efficient binary salp swarm algorithm with crossover scheme for feature selection problems. Knowl Based Syst 154:43–67

Ghadimi N, Akbarimajd A, Shayeghi H, Abedinia O (2018) Two stage forecast engine with feature selection technique and improved meta-heuristic algorithm for electricity load forecasting. Energy 161:130–142

Gheyas IA, Smith LS (2010) Feature subset selection in large dimensionality domains. Pattern Recognit 43(1):5–13

Gu S, Cheng R, Jin Y (2018) Feature selection for high-dimensional classification using a competitive swarm optimizer. Soft Comput 22(3):811–822

Gutlein M, Frank E, Hall M, Karwath A (2009) Large-scale attribute selection using wrappers. In: IEEE Symposium on

computational intelligence and data mining, 2009. CIDM'09. pp 332–339

Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3(Mar):1157–1182

Hancer E, Xue B, Karaboga D, Zhang M (2015) A binary ABC algorithm based on advanced similarity scheme for feature selection. Appl Soft Comput 36:334–348

Hancer E, Xue B, Zhang M, Karaboga D, Akay B (2018) Pareto front feature selection based on artificial bee colony optimization. Inf Sci 422:462–479

Jayaraman V, Sultana HP (2019) Artificial gravitational cuckoo search algorithm along with particle bee optimized associative memory neural network for feature selection in heart disease classification. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/s12652-019-01193-6

Kennedy J, Eberhart RC (1995) Particle swarm optimization. In: Proceedings of the IEEE international conference on neural networks, Perth, Australia, vol 4. pp 1942–1948

Kıran MS, Özceylan E, Gündüz M, Paksoy T (2012) A novel hybrid approach based on particle swarm optimization and ant colony algorithm to forecast energy demand of Turkey. Energy Convers Manag 53(1):75–83

Kohavi R, John GH (1997) Wrappers for feature subset selection. Artif intell 97(1–2):273–324

Labani M, Moradi P, Jalili M, Yu X (2017) An evolutionary based multi-objective filter approach for feature selection. In: IEEE 2017 World congress on computing and communication technologies (WCCCT), pp 151–154. https://doi.org/10.1109/WCCCT.2016.44

Lal KN, Chapelle O, Weston J, Elisseeff A (2006) Embedded methods. In: Feature extraction, foundations and applications, studies in fuzziness and soft computing, vol 207, chap 5. Springer, Berlin, p 167. https://doi.org/10.1007/978-3-540-35488-8_6

Lane MC, Xue B, Liu I, Zhang M (2013) Particle swarm optimisation and statistical clustering for feature selection. In: Australasian joint conference on artificial intelligence. Springer, Cham, pp 214–220. https://doi.org/10.1007/978-3-319-03680-9_23

Lane MC, Xue B, Liu I, Zhang M (2014) Gaussian based particle swarm optimisation and statistical clustering for feature selection. In: European conference on evolutionary computation in combinatorial optimization. Springer, Berlin, pp 133–144. https://doi.org/10.1007/978-3-662-44320-0_12

Langley P (1994) Selection of relevant features in machine learning. In: Proceedings of the AAAI fall symposium on relevance, vol 184. pp 245–271

Liu H, Yu L (2005) Toward integrating feature selection algorithms for classification and clustering. IEEE Trans Knowl Data Eng 17(4):491–502

Mafarja M, Mirjalili S (2018) Whale optimization approaches for wrapper feature selection. Appl Soft Comput 62:441–453

Mafarja M, Aljarah I, Heidari AA, Hammouri AI, Faris H, Ala'M AZ, Mirjalili S (2018) Evolutionary population dynamics and grasshopper optimization approaches for feature selection problems. Knowl Based Syst 145:25–45

Malakar S, Ghosh M, Bhowmik S, Sarkar R, Nasipuri M (2019) A GA based hierarchical feature selection approach for handwritten word recognition. Neural Comput Appl 1–20. https://doi.org/10.1007/s00521-018-3937-8

Marill T, Green D (1963) On the effectiveness of receptors in recognition systems. IEEE Trans Inf Theory 9(1):11–17

Moradi P, Gholampour M (2016) A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy. Appl Soft Comput 43:117–130

Moslehi F, Haeri A (2019) A novel hybrid wrapper–filter approach based on genetic algorithm, particle swarm optimization for feature subset selection. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/s12652-019-01364-5

Nemati S, Basiri ME, Ghasem-Aghaee N, Aghdam MH (2009) A novel ACO-GA hybrid algorithm for feature selection in protein function prediction. Expert Syst Appl 36(10):12086–12094

Pashaei E, Aydin N (2017) Binary black hole algorithm for feature selection and classification on biological data. Appl Soft Comput 56:94–106

Rao H, Shi X, Rodrigue AK, Feng J, Xia Y, Elhoseny M, Gu L (2019) Feature selection based on artificial bee colony and gradient boosting decision tree. Appl Soft Comput 74:634–642

Sharif M, Tanvir U, Munir EU, Khan MA, Yasmin M (2018) Brain tumor segmentation and classification by improved binomial thresholding and multi-features selection. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/s12652-018-1075-x

Tabakhi S, Moradi P (2015) Relevance–redundancy feature selection based on ant colony optimization. Pattern Recognit 48:2798–2811

Tao Z, Huiling L, Wenwen W, Xia Y (2019) GA-SVM based feature selection and parameter optimization in hospitalization expense modeling. Appl Soft Comput 75:323–332

Unler A, Murat A, Chinnam RB (2011) mr2PSO: a maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification. Inf Sci 181(20):4625–4641

Vergara JR, Estévez PA (2014) A review of feature selection methods based on mutual information. Neural Comput Appl 24(1):175–186

Whitney AW (1971) A direct method of nonparametric measurement selection. IEEE Trans Comput 100(9):1100–1103

Xue B (2013) Particle swarm optimisation for feature selection in classification, Victoria University of Wellington, School of Engineering and Computer Science, Ph.D. thesis

Xue B, Zhang M, Browne WN (2012) New fitness functions in binary particle swarm optimisation for feature selection. In: 2012 IEEE congress on evolutionary computation (CEC), pp 1–8. https://doi.org/10.1109/CEC.2012.6256617

Xue B, Zhang M, Browne WN (2013) Particle swarm optimization for feature selection in classification: a multi-objective approach. IEEE Trans Cybern 43(6):1656–1671

Yang W, Li D, Zhu L (2011) An improved genetic algorithm for optimal feature subset selection from multi-character feature set. Expert Syst Appl 38(3):2733–2740

Yang L, Li K, Zhang W, Zheng L, Ke Z, Qi Y (2017) Optimization of classification algorithm based on gene expression programming. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/s12652-017-0563-8

Yao C, Liu YF, Jiang B, Han J, Han J (2017) LLE score: a new filter-based unsupervised feature selection method based on nonlinear manifold embedding and its application to image recognition. IEEE Trans Image Process 26(11):5257–5269

Zeng XJ, Tao J, Zhang P, Pan H, Wang YY (2011) Reactive power optimization of wind farm based on improved genetic algorithm. In: 2nd international conference on advances in energy engineering, vol 14, pp 1362–1367