A Hybrid Multi-objective Optimization Algorithm with Improved Neighborhood Rough Sets for Feature Selection

Tao Li, Jiucheng Xu, Meng Yuan, Zhigang Gao

A Hybrid Multi-objective Optimization Algorithm with Improved Neighborhood Rough Sets for Feature Selection *

Tao Li[1,2,3][0000−0003−4032−6980], Jiucheng Xu[1,2,3][0000−0002−3860−3662], Meng Yuan[1][0000−0002−6328−4413], and Zhigang Gao[1][0000−0003−3760−357X]

[1] College of Computer and Information Engineering, Henan Normal University, Xinxiang 453007, China
[2] Key Laboratory of Artificial Intelligence and Personalized Learning in Education, Henan province, China
[3] Engineering Laboratory of Intelligence Business & Internet of Things, Henan Province, China
litao@htu.edu.cn

Abstract. Feature selection is an effective method for dimensionality reduction in machine learning and data mining. However, it is challenging to select the optimal feature subset with smaller size and higher classification accuracy from high-dimensional data. In this paper, a new approach for feature selection using multi-objective optimization algorithm with improved neighborhood rough sets is proposed. Firstly, the improved neighborhood positive region considering the classification information in the boundary domain is presented to measure the importance of feature more accurately. Then, two optimization objectives are designed to evaluate the quality of the candidate feature subsets. The non-dominated sorting operator and the crowding distance operator are employed to obtain the optimal solution sets. Finally, we utilize the feature kernel to study the relationship between the solutions in the same Pareto front. The performance of the proposed algorithm is examined on ten benchmark data sets and the results are compared with state-of-the-art algorithms to verify the validity. Experimental results show that the proposed algorithm can obtain the high-quality tradeoff between feature subset size and classification accuracy.

Keywords: Feature selection · Neighborhood rough set · Multi-objective optimization.

1 Introduction

In the are of big data, redundant features and irrelevant features in high dimensional data may lead to the curse of dimensionality, which presents a challenge

* Supported by the National Natural Science Foundation of China under Grant 61976082, the Key Scientific Research Project of Henan Provincial Higher Education under Grant 22B520013 and the Doctoral Scientific Research Foundation of Henan Normal University under Grant 20210248.

for machine learning and data ming [1]. Dimensionality reduction technology can address the issues powerfully. It mainly consists of feature selection (FS) method and feature extraction (FE) method [2]. Although FE can reduce the dimensionality, it may produce new features. While FS can obtain a feature subset that maintains the physical meanings of original features. More importantly, it can enhance the performance of the training model. So, we focus mainly on feature selection in this paper.

In recent years, many researchers have studied extensively the feature selection methods based on evolutionary algorithms (EAs) [3–5]. From the perspective of optimizing the number of objectives, evolutionary algorithm is divided into single objective optimization and multi-objective optimization. Most feature selection methods based on single-objective evolutionary algorithms aim to maximize the classification accuracy or minimize the number of features. While feature selection based on multi-objective evolutionary algorithms have attracted much attention by researchers. But most of them only integrate two objectives or more into a single objective without considering the relationship between objectives. For example, Consider the relationship between feature number and classification accuracy, Jimnez F. et al. [6] proposed a feature selection methodology composed by the application of the multi-objective evolutionary algorithm for online sales forecasting. Wang Z. [7] presented a multi-objective evolutionary algorithm with class-dependent redundancy for feature selection based on a relevance measure and new redundancy measure. It is generally known that multi-objective optimization problems are expert in solving conflicts between objectives. Some papers apply non-dominated relationship to tackle the conflicts between classification performance and the size of feature subset. For instance, Xue B. et al. [8] presented the study on multi-objective particle swarm optimization for feature selection and generates a Pareto front of non-dominated solutions (feature subsets). For tackling the feature selection problem with unreliable data, Zhang Y. et al. [9] proposed an effective multi-objective feature selection algorithm based on bare-bones particle swarm optimization, where the reliability and the classification accuracy are taken as the two objectives. Xu H. [10] proposed a duplication analysis-based EA (DAEA) for multiobjective feature selection, and obtained the good classification and generalization results.

Although many EAs have been successfully applied to feature selection problem, there are still two issues that need further study. Firstly, the information contained in the classification boundary region is not been considered. It is clear that the redundant features and relevant features can convert to each other. This phenomenon directly affects the fault tolerance of classification systems. So, we should take the boundary information into account to enhance the relationships between the condition features and decision features so as to further improve the performance of EAs. Secondly, most EAs only present the solutions without relationship between different solutions is analyzed. Many solutions can be obtained by multi-objective evolution algorithm, but it is still a thorny issue to select the most reasonable solution from solution sets. Therefore, we need to further study the relationship between the solutions in the same Pareto front.

Motivated by the above two main issues, we present a hybrid multi-objective optimization algorithm with improved neighborhood rough sets for feature selection. The performance of proposed method is examined on ten publicly available data sets and twelve algorithms to verify the effectiveness. The remaining part of the paper is organized as follow: Section 2 provides the related work. In Section 3, the proposed algorithm is presented. Section 4 gives the experimental design and result analysis. Finally, the conclusion and the future research direction are presented in Section 5.

2 Related Work

2.1 Neighborhood Rough Set

Feature selection based on neighborhood rough set is to find a minimal feature subset with the same distinguishing ability as all features. While the ability to distinguish is measured by the size of positive region. In other words, we tend to find the lager positive region in which the significant features are essential for classification. Suppose the universal $U = \{x_1, x_2, \cdots, x_{|U|}\}$, attribute set $A = \{a_1, a_2, \cdots, a_{|A|}\}$, $A = C \cup D$ and $C \cap D = \oslash$, where C is the condition attributes and D is the decision attributes, the value domain $V = V_C \cup V_D$, V_C and V_D is the value of the C and D, information function $F : U \times A \to V$, then $S = (U, A, V, F)$ is denoted as an information system.

Definition 1. Information system $S = (U, A, V, F)$, for $\forall B \subseteq A$, the undistinguishable relation of B on the domain U is defined as

$$IND(B) = \{(x, y) \in U \times U | f(x, a) = f(y, a), \forall a \in B\} \tag{1}$$

The partition of the universal U denoted as $U/IND(B) = \{[x]_B | x \in U\}$, where $[x]_B$ represents the equivalent class of any object x in U under attribute set B. Here the $[x]_B = \{y \in U | (x, y) \in IND(B)\}$ is called knowledge granularity.

Definition 2. Suppose a neighborhood decision system $NDS = \{U, A, V, F, \varepsilon\}$, $A = C \cup D$ and $\forall B \subseteq C$, for $\forall x_i \in U$, the neighborhood of the x_i is $\varepsilon(x_i) = \{x | \triangle(x, x_i) \leq \varepsilon, x \in U\}$. The decision attribute D divides U into N equivalent classes $\{X_1, X_2, \cdots, X_N\}$. So the neighborhood negative region and positive region of decision D with respect to B are respectively denoted as

$$\overline{N_B} X = \bigcup_{i=1}^{N} \{x_i | \varepsilon_B(x_i) \cap X \neq \oslash, x_i \in U\} \tag{2}$$

$$\underline{N_B} X = \bigcup_{i=1}^{N} \{x_i | \varepsilon_B(x_i) \subseteq X, x_i \in U\} \tag{3}$$

It is clear that the boundary region $BN_B(X) = \overline{N_B} X - \underline{N_B} X$. The importance of attributes is measured by $|\underline{N_B} X|/|U|$, where the $|\underline{N_B} X|$ is the number of sample contained by X within a certain neighborhood and the $|U|$ is the size of sample space. Hence, we can see that the hidden information in the boundary area is not considered when calculating the importance of attributes.

2.2 Multi-objective Optimization

Considering the global optimization ability, the multi-objective evolutionary algorithm based on population search is suitable for solving the feature selection problem [11]. In the process of feature selection, the candidate feature subset usually contains relevant and redundant features, and the redundant features should be removed to reduce the size of the feature subset, so the size of feature subset and the classification performance are considered as two optimization objectives. In order to judge the pros and cons of the solution of different objectives, the definition of the Pareto-dominance and the Pareto front are described [12, 13].

Definition 3. Suppose objective vector $\boldsymbol{y} = (f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \cdots, f_n(\boldsymbol{x}))$, where solution vector $\boldsymbol{x} = (x_1, x_2, \cdots, x_m)$. The Ψ is the solution space, for $\forall a \in \Psi$, $\forall b \in \Psi$, if $\forall i \in 1, 2, \cdots, n$ makes $f_i(a) \leq f_i(b)$ and $\exists j \in 1, 2, \cdots, n$ makes $f_j(a) < f_j(b)$, then it is called a dominates b and denoted as $a \Rightarrow b$.

Definition 4. For decision variables $\boldsymbol{x} \in \boldsymbol{R}^m$, if it does not exist $\boldsymbol{c} \in \boldsymbol{R}^m$ makes $\boldsymbol{c} \Rightarrow \boldsymbol{x}$, then $\boldsymbol{x}$ denote as the Pareto optimal solution and $T^\star$ is the Pareto optimum solutions set, so we defined the Pareto front (PF) as:

$$PF = \{F(\boldsymbol{x}) = (f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \cdots, f_n(\boldsymbol{x})) | \boldsymbol{x} \in T^\star\} \tag{4}$$

3 Proposed Multi-objective Optimization Algorithm with Improved Neighborhood Rough Sets

3.1 Improved Neighborhood Rough Sets

Here, the neighborhood rough set is adopted to construct granular models for feature selection. The neighborhood rough set theory holds that the more detailed description of attributes, the higher the division between samples. But there are still two defects: first, the sample in the positive region is calculated without considering the information in the boundary area, which leads to the lack of hidden classification information. Second, there are still redundant attributes in condition attributes, and it is disadvantageous for obtaining low-dimensional feature subset. In order to solve the issues, a new neighborhood positive region computing method is proposed in this paper.

Definition 5. Given neighborhood decision system $NDS = \{U, A, V, F, \varepsilon\}$, $A = C \cup D$, C is the condition attribute and D is the decision attribute. The decision attribute D divides U into K equivalent classes $X_1, X_2, \cdots, X_K$, and the conditional attribute C divided U into $\{C_1, C_2, \cdots, C_n\}$ based on the neighborhood. Then the new neighborhood positive region and the dependence degree of decision attributes on conditional attributes are respectively defined as

$$\underline{LN_C}X = \bigcup_{i=1}^{N} \{x_i | argmax(\varepsilon_C(x_i) \cap X), x_i \in U\} \tag{5}$$

$$\varphi_C(D) = |\underline{LN_C}X|/|U| \tag{6}$$

The value of $\varphi_C(D)$ represents the rate of the sample being divided in the positive region. The greater value of the $\varphi_C(D)$ indicates that the partition under the attributes (feature subsets) can obtain stronger classification ability. While the smaller value of the $\varphi_C(D)$, the lower classification ability under the attributes.

Theorem 1. $NDS = \{U, A, V, F, \varepsilon\}$, the $\triangle$ is the measure function on U, if A_1, $A_2 \subseteq C$ and $A_1 \subseteq A_2$, then $\forall X \subseteq U$, $\underline{LN_{A_1}}X \subseteq \underline{LN_{A_2}}X$.

Proof. Suppose $x_i \in U$ and $A_1 \subseteq A_2$, if $x_i \in \varepsilon_{A_2}$ we can get that $x_i \in \varepsilon_{A_1}$ according to $\triangle_{A_1}(x_1, x) \leq \triangle_{A_2}(x_1, x)$. So we have $\varepsilon_{A_1}(x) \supseteq \varepsilon_{A_2}(x)$. Given $\varepsilon_{A_1}(x) \subseteq \underline{LN_{A_1}}X$, where X is the samples of a decision class and it is easy to obtain $\varepsilon_{A_2}(x) \subseteq \underline{LN_{A_2}}X$. While there may exists x_i that makes $\varepsilon_{A_1}(x) \nsubseteq \underline{LN_{A_1}}X$ and $\varepsilon_{A_2}(x) \subseteq \underline{LN_{A_2}}X$, so we get $\underline{LN_{A_1}}X \subseteq \underline{LN_{A_2}}X$.

Inference 1. If $A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots \subseteq C$, then it can get $\varphi_{A_1}(D) \leq \varphi_{A_2}(D) \leq \varphi_{A_3}(D) \leq \cdots \leq \varphi_C(D)$.

The inference 1 shows that the more the selected attributes, the more accurate the description of the sample. In other words, the sample in the different classes will have a greater difference with more features, which helps to distinguish the samples. But the attribute subset often have irrelevant attributes and redundant attributes that may increase the complexity of classification models. Therefore, it is necessary to reduce the size of subset of candidate features.

Definition 6. $NDS = \{U, A, V, F, \varepsilon\}$, $A = C \cup D$ and $E \subseteq C$, If E meets the two conditions simultaneously : 1) $\varphi_E(D) = \varphi_C(D)$; 2) $\forall a \in E$, $\varphi_E(D) > \varphi_{E-\{a\}}(D)$, then we call attribute subset E is a relative reduction of C.

Definition 6 explains that the attribute subset after reduction maintains the core attributes of knowledge system. This helps to establish an efficient classification model, because it removes the irrelevant features and redundant features.

3.2 Design of Two Objective Functions

As mentioned above, the combination of different features directly affects classification performance. That is to say, when the important features are deleted, it deteriorates the classification performance. While adding the important features it can improve the classification performance. Obviously, there is a conflict between the number of important features and the classification performance. It is worth noting that the measurement of the importance of features has been calculated by Eq.(6). In order to evaluate each candidate feature subset better, objective functions F_1 and F_2 are designed. Among them, F_1 is the feature selection rate and it is used to evaluate the size of feature subset on different scale data sets. while F_2 is the classification error rate. Therefore, the two objective functions are respectively defined as

$$F_1 = SR = |N_s|/|N_f| \tag{7}$$

$$F_2 = Err = \frac{FP + FN}{FP + FN + TP + TN} \tag{8}$$

In Eq.(7), the $|N_s|$ and the $|N_f|$ stand for the number of selected features and original features, respectively. While in Eq.(8), the FP, FN, TP and TN denote false positives, false negatives, true positives and true negatives, respectively.

3.3 Feature Kernel Set for Pareto Front

To further study the relationship between the solutions in the same Pareto front, the intersection of feature subsets in the Pareto front solution set is calculated. The purpose of this process is to obtain the key features shared by different candidate feature subsets, which are the essential features in the non-dominated solutions. Suppose $S = \{S_1, S_2, \cdots, S_n\}$ is the feature subsets in the PF respects to F_1 and F_2, where n is the number of solutions. Then, the feature kernel set (FKS) can be defined as $FKS = \bigcap_{i=1}^{n} S_i$.

3.4 Complete Procedure of the Proposed MOINR

In this chapter, a new approach to feature selection using multi-objective optimization algorithm with improved neighborhood rough sets is presented. The proposed algorithm is illustrated in Algorithm 1. The code of lines 5-10 explain the feature importance calculated by the new dependence degree of decision attributes on conditional attributes. The lines 18-19 describe the measure value of each candidate solution on two objective functions. While the lines 21 is to calculate the non-dominant solutions and constructing non-dominant Front. In addition, the individual crossover operator, individual mutation operator and population update operator are listed in lines 23-26.

Assuming that the number of objective functions is N_o and the size of population is N_p. At the same time, the number of the initial feature and the selected feature are recorded as N_f and N_s, respectively. It is easy to know that the time complexity of NSO is $O(N_o \times N_p^2)$, and the time complexity of $ICDO$ is $O(N_o \times N_p \times logN_p)$, while the time complexity of computing neighborhood positive domains is $O(N_f \times N_s^2 \times N_p \times logN_p)$. In this case, the time complexity of $MONPR$ is calculated as $O(T \times (N_f \times N_s^2 \times N_p \times logN_p + N_o \times N_p^2 + N_o \times N_p \times logN_p))$. According the asymptotic time complexity theory, the time complexity of $MONPR$ is reduce to $O(T \times N_s^3 \times N_p \times logN_p)$ at least.

4 Experimental Design and Result Analysis

4.1 Data Sets and Parameter Setup

The simulations are conducted on Core(TM) i5-4440, 3.10 GHz CPU, 8 GB RAM and the proposed algorithm are implemented in MATLAB R2014a and WEKA 3.8.0 software. Table 1 shows that the ten data sets are adopted for testing the proposed algorithm, and these data sets are available from the University of California Irvine (UCI).

Algorithm 1 Multi-objective Optimization Algorithm with Improved Neighborhood Rough Sets (MOINR)

Input: $Data = (x_1, x_2, \cdots, x_N, y)$, the number of iteration T, the population size N_p and feature importance lower limit λ, neighborhood value ε.
Output: Pareto solution sets

Initialization population individual.
$CS = \oslash, t = 0$ //Initialization candidate subset and iteration number
while the t meet maximum iteration T
for i =1 to N_p **do**
  $F_{set}(i) = Data(pop_{index}(i) == 1)$ // Select the features of the index value equal to 1
  for j =1 to $|F_{set}|$ **do**
    $A_j \in F_{set} - CS$
    $\underline{LN_{A_j \cup CS}} X = \bigcup_{k=1}^{N} \{x_k | argmax(\varepsilon_{A_j \cup CS}(x_k) \cap X), x_k \in U\}$
    $\varphi_{A_j \cup CS}(D) = |\underline{LN_{A_j \cup CS}} X|/|U|$
    $INV(A_j) = \varphi_{A_j \cup CS}(D) - \varphi_{A_j}(D)$
  end for
  if $INV(A_j, CS) < \lambda$ **then**
    $CS = CS \cup A_j$
    goto 6
  else
    $OFS(i) = CS$
  end if
  $indiv_1 = F1(OFS)$ //Calculating objective functions F1.
  $indiv_2 = F2(OFS)$ //Calculating objective functions F2.
end for
$PF = NSO(pop)$ // Measure the non-dominating relationship of each individual.
$pop_{distance} = ICDO(pop, front)$ // Calculating the distance between individuals in the front
$Chrom(pop_c) = Chrom(pop)$ // Individual crossover operation.
$Chrom(pop_m) = Chrom(pop_c)$ // Individual mutation operation.
$Pop = pop \cup pop_m$ // Merge the father population and offspring population.
$Pop = newsort(Pop)$ //Form new population with the same size.
$Front = update(front)$
$t = t + 1$
end While

In order to verify the credibility and stability of the experimental results, the process of randomly selection is repeated ten independent times for obtaining the statistically meaningful experimental results and the artificial neural network is adopted as a classifier. In the process of iteration, the crossover operator and mutation operator are used for produce new population. The crossover factor $CF = 0.8$ and mutation factor are $MF = 0.1$. In order to ensure that each feature is selected with the same probability, the length of the P_1 is equal to the dimension of the original dataset. While the population size P_n=20 and the number of iteration T=100. In the process of calculating the importance of features, the neighborhood value ε is fixed to 0.9 and the threshold of feature importance λ is set as 0.01.

Table 1: Experimental data sets description

No.	Data sets	No. of features	No. of instances	No. of classes
1	Wine	13	178	3
2	Vehicle	18	846	4
3	Lymph	18	148	4
4	WDBC	30	569	2
5	Ionophere	34	351	2
6	SPECTE	44	80	2
7	Sonar	60	208	2
8	Synthetic	60	600	6
9	Hill-valley	100	606	2
10	Musk	166	476	2

In Table 2, FKS and FS respectively represent the feature kernel set and the best feature subset in Pareto front obtained by MOINR . While the ACC is the classification accuracy of FS. It can be seen that the number of feature contained by FKS is not more than 5 on eight data sets. When the FKS is combined with other features, it can significantly enhance classification ability of the algorithm. For example, when the KFS of Wine data set combined with the 10th and 11th features, its classification accuracy can reach to 97.58%. For WDBC data set, 98.73% of the classification accuracy was obtained when the 18th and 29th features are added to the corresponding KF. Therefore, the space of features combination is reduced by the obtained FKS. More importantly, it is helpful to select smaller subset with higher classification accuracy.

4.2 Performance Comparison Between MOINR and Other Algorithms

In order to verify the effectiveness and advantages of the proposed algorithm, the performance of the MOINR is evaluated by comparing with three groups representative methods. 1) Classical filter algorithms: ReliefF [14] , mRMR [15], FSIG [16] and SBMLR [17]. 2) Single objective wrapper algorithms: BGAFS [18], MDEFS [19], BPSO [20] and BACO [21]. 3) Multi-objective wrapper algorithms: MPPSO [22], MOEA/D [23] , FWSP [24] and NSGAII [12] MORS.

Table 2: The best classification accuracy by feature kernel sets in MOINR

Dataset	Indicator	ID of Features
Wine	FKS	1,12,13
	FS	1,10,11,12,13
	ACC	**97.58%**
Vehicle	FKS	1,10
	FS	1,7,10,13,14
	ACC	**72.22%**
Lymph	FKS	2,13,18
	FS	2,7,9,13,15,18
	ACC	**88.31%**
WDBC	FKS	1,16,22,25
	FS	1,16,18,22,25,29
	ACC	**98.73%**
Ionophere	FKS	10,17
	FS	3,4,9,10,17
	ACC	**91.83%**
SPECTE	FKS	19,40
	FS	4,17,18,19,21,22,25,26,27,29,30,33,36,38,39,40,42
	ACC	**80.31%**
Sonar	FKS	11,36,47
	FS	11,22,36,47
	ACC	**83.99%**
Synthetic	FKS	40,43,56
	FS	19,27,35,40,43,46,57
	ACC	**95.33%**
Hill-valley	FKS	1,4,23,59,69,70,75,82,98
	FS	1,4,16,19,23,24,43,45,47,51,56,59,69,70,75,76,82,98
	ACC	**76.11%**
Musk	FKS	9,10,28,44,49,53,56,71,72,79,83,99,107,132,135,138,147,152,161
	FS	9,10,17,20,28,44,46,49,53,56,64,66,71,72,77,78,79,82,83,85,87,90, 94,95,99,105,107,111,117,120,125,131,132,135,138,147,149,152,156,161
	ACC	**81.19%**

In this paper, the classification accuracy (ACC) and selection rate (SR) are taken as two indicators to evaluate the performance of the proposed algorithm. In order to show the improvement of classification performance after reduction, we also present the results using the original data set (OD). In the tables, the W/L/T represents the number of wins/loss/ties for MOINR in comparison with the other algorithms on the ten data sets. The BAVE and BSTD represent the average and standard deviation of the all data sets for each algorithm, respectively. While the AVE and STD show the average and standard deviation of the ACC and SR for each data set in ten runs, respectively. In addition, we rank the algorithms according to the BAVE and the values are shown in the parentheses. The maximum ACC and minimum SR are highlighted in bold for each data set. Besides, the T-test is utilized to test the significance of the proposed algorithm.

A. Comparison with classical filter algorithms

Table 3 and 4 show the average classification accuracy and average feature selection rate. In the tables, four representative classical filter algorithms are compared with the proposed algorithm. 1) ReliefF is a filtering algorithm based on sample learning that can select the appropriate feature according the threshold of feature importance. 2) mRMR is developed based on the redundancy of features

and the correlation between features and labels, which used mutual information as the feature selection criterion. 3) FSIG is an effective feature selection method based on information gain, whose selection criterion is the contribution of the feature to the classification system. 4) SBMLR is a sparse multinomial logistic regression method that can obtain the most informative features.

Table 3: Average ACC to MOINR and the comparison with four classical filter algorithms in ten runs

Dataset	OD		MOINR		ReliefF		mRMR		FSIG		SBMLR	
	AVE(%)	STD(%)	AVE(%)	STD(%)	AVE(%)	STD(%)	AVE(%)	STD(%)	AVE(%)	STD(%)	AVE(%)	STD(%)
Wine	96.63	0.00	**94.36**	7.43	87.36	12.38	91.64	7.13	92.42	6.65	91.50	5.27
Vehicle	44.80	0.00	**72.22**	8.78	41.91	3.64	42.59	1.38	41.42	1.46	45.57	4.28
Lymph	82.43	0.00	**86.62**	4.90	79.28	3.88	78.68	5.79	77.55	4.39	75.08	3.14
WDBC	89.86	0.00	**98.51**	0.68	92.37	2.01	88.99	2.81	91.06	1.56	85.14	1.54
Ionophere	82.34	0.00	**92.60**	3.27	87.48	4.53	89.03	2.77	86.56	2.06	80.77	6.87
SPECTE	76.25	0.00	**80.31**	8.34	79.43	1.92	77.25	2.42	81.08	2.26	74.25	10.02
Sonar	67.79	0.00	**86.45**	2.45	69.64	2.48	69.04	2.73	68.40	2.88	72.48	4.00
Synthetic	94.67	0.00	**95.45**	2.00	53.33	1.91	74.48	5.36	75.67	4.90	86.16	8.88
Hill-valley	51.98	0.00	**81.05**	0.66	56.27	8.00	51.82	0.50	51.67	0.43	52.06	0.12
Musk	75.21	0.00	**93.06**	2.36	67.06	5.05	57.21	2.59	67.97	5.00	69.90	2.46
W/L/T	10/0/0		0/0/10		10/0/0		10/0/0		10/0/0		10/0/0	
T-test	0.0074		—		0.0038		0.0028		0.0029		0.0005	

Table 4: Average SR to MOINR and the comparison with four classical filter algorithms in ten runs

Dataset	OD		MOINR		ReliefF		mRMR		FSIG		SBMLR	
	AVE(%)	STD(%)	AVE(%)	STD(%)	AVE(%)	STD(%)	AVE(%)	STD(%)	AVE(%)	STD(%)	AVE(%)	STD(%)
Wine	100	0.00	**32.05**	17.82	36.92	14.39	38.46	19.86	36.92	14.39	34.62	18.84
Vehicle	100	0.00	**27.68**	15.21	27.78	15.21	27.78	14.34	27.78	15.21	27.78	15.21
Lymph	100	0.00	**26.36**	23.75	27.78	15.21	27.78	14.34	27.78	15.21	27.78	15.21
WDBC	100	0.00	29.52	10.44	26.67	14.91	18.33	10.09	26.67	14.91	**5.00**	2.40
Ionophere	100	0.00	23.95	15.23	26.47	14.85	16.18	8.90	26.47	14.85	**7.35**	4.00
SPECTE	100	0.00	36.11	16.67	26.14	14.76	12.50	6.88	18.18	10.16	**6.80**	3.50
Sonar	100	0.00	20.00	2.64	17.50	9.86	9.17	5.05	13.33	7.45	**7.50**	4.10
Synthetic	100	0.00	29.00	9.47	17.50	9.86	**8.33**	4.30	13.33	7.45	13.33	7.50
Hill-valley	100	0.00	32.43	5.90	10.50	5.92	**3.00**	1.41	8.00	4.47	3.50	0.71
Musk	100	0.00	27.11	2.92	48.19	2.69	**3.01**	1.56	4.82	2.69	3.31	1.82
W/L/T	10/0/0		0/0/10		4/6/0		3/7/0		4/6/0		3/7/0	
T-test	0.0000		—		0.6151		0.0128		0.0451		0.0045	

The results of Table 3 indicate that the MOINR can obtain the highest classification accuracy in ten runs. The reason is that the four comparison algorithms belong to the filter type, and the evaluation of feature importance only according to the relationship between the features and the classification labels. While in the proposed algorithm, the relationship among features and the relationship between features and classification labels are considered simultaneously to measure

the quality of the feature subset. Therefore, the proposed MOINR can improve the classification accuracy. From Table 4, it seems that the MOINR algorithm is not outstanding in the feature selection rate, which may be caused by the larger threshold of feature importance.

B. Comparison with single objective wrapper algorithms

It is meaningful to compare the proposed algorithm with the single objective wrapper algorithm, because the multi-objective method can better explain how to find the best compromise solutions. Hence, four single objective algorithms are employed in this paper. The BGAFS and the MDEFS are binary genetic algorithm and binary differential evolution algorithm, respectively. The objective functions of the BGAFS and MDEFS are the ratio of within-class distance and between-class distance. The smaller the ratio, and the better the quality of the selected feature subset. While the BPSO and the BACO are the binary particle swarm optimization algorithm and binary ant colony optimization algorithm, respectively. The objective functions of BPSO and BACO are the classification accuracy of the selected feature subset on the training model.

Table 5: Average ACC to MOINR and the comparison with four single objective wrapper algorithms in ten runs.

Dataset	OD		MOINR		BGAFS		BPSO		MDEFS		BACO	
	AVE(%)	STD(%)	AVE(%)	STD(%)	AVE(%)	STD(%)	AVE(%)	STD(%)	AVE(%)	STD(%)	AVE(%)	STD(%)
Wine	96.63	0	94.36	7.43	94.94	1.38	**98.25**	1.16	96.07	1.21	97.33	1.65
Vehicle	44.80	0	**72.22**	8.78	48.88	0.92	52.29	3.81	48.88	0.92	49.75	4.89
Lymph	82.43	0	**86.62**	4.90	82.43	0.00	50.35	5.68	82.43	0.00	89.42	1.86
WDBC	89.86	0	**98.51**	0.68	92.75	1.57	98.49	0.48	90.76	1.49	98.40	0.30
Ionophere	82.34	0	92.60	3.27	89.60	38.24	**93.55**	0.38	83.83	14.30	94.85	0.40
SPECTE	76.25	0	**80.31**	8.34	78.75	3.06	69.65	2.93	76.56	3.59	59.67	5.70
Sonar	67.79	0	86.45	2.45	67.55	3.65	88.55	1.44	63.58	7.19	**89.34**	0.70
Synthetic	94.67	0	**95.45**	2.00	77.67	0.00	75.02	5.42	77.67	0.00	71.73	3.50
Hill-valley	51.98	0	**81.05**	0.66	51.94	0.07	78.81	0.74	51.90	0.10	78.10	0.50
Musk	75.21	0	**93.06**	2.36	70.94	0.60	87.73	1.20	75.84	0.00	86.97	0.90
W/L/T	10/0/0		0/0/10		10/0/0		7/3/0		9/1/0		6/4/0	
T-test	0.0074		—		0.0052		0.0613		0.0025		0.1015	

Table 5 reveals the average classification accuracy of comparative approaches in ten runs. We can see that the proposed algorithm can get the highest average classification accuracy on seven data sets. In most cases, the proposed algoorithm also has a significant advantage over BGAFS, MDEFS, BPSO and BACO. This is because the multi-objective algorithm can select the Pareto solution sets instead of the single solution obtained by the single objective algorithm. It helps us to find more reasonable and comprehensive feature subsets. In addition, Table 6 reflects the average feature selection rate obtained by MOINR against other four single algorithms. It can be observed that the MOINR has gained the minimum average feature selection rate on half of the data sets.

Table 6: Average SR to MOINR and the comparison with four single objective wrapper algorithms in ten runs.

Dataset	OD		MOINR		BGAFS		BPSO		MDEFS		BACO	
	AVE(%)	STD(%)	AVE(%)	STD(%)	AVE(%)	STD(%)	AVE(%)	STD(%)	AVE(%)	STD(%)	AVE(%)	STD(%)
Wine	100	0	**32.05**	17.82	40.38	3.85	46.15	0.00	96.15	4.44	42.30	9.93
Vehicle	100	0	27.68	15.21	38.89	0.00	**26.39**	5.32	38.89	0.00	30.55	7.17
Lymph	100	0	**26.36**	23.75	38.89	0.00	27.78	5.56	38.89	0.00	30.55	7.17
WDBC	100	0	**29.52**	10.44	42.50	1.67	31.33	4.30	31.67	1.92	18.33	4.30
Ionophere	100	0	**23.95**	15.23	38.24	0.00	26.18	3.80	24.26	6.52	16.18	3.80
SPECTE	100	0	36.11	16.67	41.48	3.40	13.43	2.82	27.84	7.04	**12.50**	2.94
Sonar	100	0	20.00	2.64	39.58	1.60	29.17	2.15	25.00	2.36	**9.17**	2.15
Synthetic	100	0	29.00	9.47	**23.33**	0.00	29.17	2.15	24.33	0.00	29.17	2.15
Hill-valley	100	0	32.43	5.90	37.00	0.00	26.75	2.20	31.00	1.41	**5.50**	1.30
Musk	100	0	**27.11**	2.92	40.96	1.59	28.16	0.57	27.71	0.00	33.31	0.78
W/L/T	10/0/0		0/0/10		9/1/0		8/2/0		7/3/0		5/5/0	
T-test	0.0000		—		0.0016		0.9924		0.2439		0.1895	

C. Comparison with multi-objective wrapper algorithm results

In order to further verify the competitiveness of the proposed algorithm, the advanced multi-objective wrapper algorithms are compared and analyzed, such as MPPSO, MOEA/D, FWSP and NSGAII. As it can be seen in the tables, the MPPSO is the multi-population based particle swarm optimization, which adopts average classification accuracies and the number of features as two optimization objectives. The MOEA/D is multi-objective evolutionary algorithm based decomposition, which also can obtain the better trade-off among the objective functions. And the NSGAII is a fast and elitist multi-objective genetic algorithm, which adopts the non-dominated sorting to analyze the relationship between different objective function solutions.

Table 7: Average ACC to MOINR and the comparison with four multi-objective wrapper algorithms in ten runs.

Dataset	OD		MOINR		MPPSO		MOEA/D		FWSP		NSGAII	
	AVE(%)	STD(%)	AVE(%)	STD(%)	AVE(%)	STD(%)	AVE(%)	STD(%)	AVE(%)	STD(%)	AVE(%)	STD(%)
Wine	96.63	0	94.36	7.43	89.15	7.96	**96.07**	9.21	94.94	6.32	83.86	8.92
Vehicle	44.80	0	**72.22**	8.78	47.13	9.48	68.50	10.44	64.80	12.78	43.67	9.36
Lymph	82.43	0	**86.62**	4.90	85.67	6.86	84.43	5.93	81.75	5.04	84.99	7.08
WDBC	89.86	0	**98.51**	0.68	97.48	2.26	96.32	7.55	94.23	10.21	95.15	8.89
Ionophere	82.34	0	92.60	3.27	**95.10**	1.33	86.50	8.82	84.33	8.82	94.74	7.19
SPECTE	76.25	0	**80.31**	8.34	62.03	7.85	74.54	5.63	79.74	8.09	72.93	6.23
Sonar	67.79	0	86.45	2.45	**91.14**	1.96	79.32	4.19	80.01	36.67	83.17	5.33
Synthetic	94.67	0	95.45	2.00	66.96	6.10	95.80	3.13	**97.00**	3.33	81.47	6.21
Hill-valley	51.98	0	**81.05**	0.66	77.09	2.39	71.67	6.76	79.68	5.42	69.04	8.84
Musk	75.21	0	**93.06**	2.36	85.80	1.13	85.43	24.10	90.92	8.61	84.27	4.26
W/L/T	10/0/0		0/0/10		8/2/0		8/2/0		9/1/0		10/0/0	
T-Test	0.0074		—		0.0498		0.0050		0.0135		0.0105	

In Table 7, we can find that MOINR can effectively reduce the average classification rate on the Vehicle, Lymph, WDBC, SPECTF, Hill-valley and Musk

Table 8: Average SR to MOINR and the comparison with four multi-objective wrapper algorithms in ten runs.

Dataset	OD		MOINR		MPPSO		MOEA/D		FWSP		NSGAII	
	AVE(%)	STD(%)	AVE(%)	STD(%)	AVE(%)	STD(%)	AVE(%)	STD(%)	AVE(%)	STD(%)	AVE(%)	STD(%)
Wine	100	0	**32.05**	17.82	43.08	24.68	37.25	13.67	40.82	14.29	47.69	14.79
Vehicle	100	0	**27.68**	15.21	45.37	25.60	38.46	44.44	31.53	17.14	35.81	19.07
Lymph	100	0	26.36	23.75	**25.00**	22.54	37.58	18.93	28.45	12.32	41.67	20.50
WDBC	100	0	**29.52**	10.44	19.63	12.30	30.17	9.35	31.74	12.96	20.83	11.23
Ionophere	100	0	**23.95**	15.23	34.41	18.91	27.91	12.47	26.16	8.82	31.99	9.61
SPECTE	100	0	36.11	16.67	24.30	11.22	**34.82**	11.75	40.27	15.06	40.96	12.85
Sonar	100	0	**20.00**	2.64	28.83	10.25	21.67	5.33	24.62	6.11	23.51	6.92
Synthetic	100	0	29.00	9.47	31.50	11.49	28.38	7.19	**28.34**	8.56	30.17	11.23
Hill-valley	100	0	32.43	5.90	37.55	12.13	35.74	9.52	35.18	10.11	**23.71**	12.31
Musk	100	0	**27.11**	2.92	34.16	5.37	30.22	7.39	31.41	6.83	32.87	6.59
W/L/T	10/0/0		0/0/10		8/2/0		8/2/0		9/1/0		8/2/0	
T-test	0.0000		—		0.2130		0.0209		0.0015		0.123	

data sets. From Table 8, it is observed that the smallest average feature selection rate achieved by MOINR on six datasets.

Furthermore, the significant test on classification accuracy and selection rate are presented in the paper. Obviously, the values of T-test are below 0.05 for most cases in the tables. It shows that the improvement of the proposed algorithm in classification accuracy and feature selection rate is notable. Thus, the proposed multi-objective optimization algorithm with improved neighborhood rough sets can accurately depict the uncertainty of samples in classification boundaries. More importantly, the MOINR can achieve the better compromise solutions.

5 Conclusion and Future Work

In this paper, a hybrid multi-objective optimization algorithm with improved neighborhood rough sets is proposed to solve the feature selection problem. The neighborhood positive region calculation method is improved to describe the discriminability of samples more accurately. Then the improved neighborhood rough model is used to select the important features. At the stage of evolutionary computation, the classification error rate and feature selection rate two objectives are designed to evaluate the candidate feature subset. Besides, the scale of feature subset is reduced and classification accuracy is improved by the feature kernel set. Therefore, the proposed algorithm can select the optimal feature subset with smaller size and higher classification accuracy.

Although multi-objective optimization algorithm can effectively handle feature selection problem, it is not excellent in terms of time cost. In the future works, we will focus on the study of faster non-dominated sorting methods and multi-objective optimization algorithms for more than two optimization objectives to deal with high-dimensional data sets.

References

1. Guyon, I., Elisseeff, A.: An introduction to variable feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (2003)
2. Chandrashekar, G., Sahin, F.: A survey on feature selection methods. Pergamon Press, Inc. (2014)
3. Faris, H., Mafarja, M.M., Heidari, A.A., Aljarah, I., Al-Zoubi, A.M., Mirjalili, S., Fujita, H.: An efficient binary salp swarm algorithm with crossover scheme for feature selection problems. Knowl. Based Syst. **154**, 43–67 (2018)
4. Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S., Coello, C.A.C.: Survey of multiobjective evolutionary algorithms for data mining: Part ii. IEEE Trans. Comput. **18**(1), 20–35 (2014)
5. Das, A.K., Das, S., Ghosh, A.: Ensemble feature selection using bi-objective genetic algorithm. Knowl. Based Syst. **123**, 116–127 (2017)
6. Jimnez, F., Snchez, G., Garca, J.M., Miralles, L., Miralles, L.: Multi-objective evolutionary feature selection for online sales forecasting. Neurocomputing **234**(C), 75–92 (2016)
7. Wang, Z., Li, M., Li, J.: A multi-objective evolutionary algorithm for feature selection based on mutual information with a new redundancy measure. Inf. Sci. **307**, 73–88 (2015)
8. Xue, B., Zhang, M., Browne, W.N.: Particle swarm optimization for feature selection in classification: A multi-objective approach. IEEE Trans. Cybern. **43**(6), 1656–1671 (2013)
9. Zhang, Y., Gong, D.W., Zhang, W.Q.: Feature selection of unreliable data using an improved multi-objective pso algorithm. Neurocomputing **171**(C), 1281–1290 (2015)
10. Xu, H., Xue, B., Zhang, M.: A duplication analysis-based evolutionary algorithm for biobjective feature selection. IEEE Transactions on Evolutionary Computation **25**(2), 205–218 (2021)
11. Jin, X., Bo, T., He, H., Hong, M.: Semisupervised feature selection based on relevance and redundancy criteria. IEEE Trans. Neural Netw. Learn. Syst. **28**(9), 1974–1984 (2017)
12. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: Nsga-ii. IEEE Trans.Evol. Comput. **6**(2), 182–197 (2002)
13. Zhu, Y., Liang, J., Chen, J., Ming, Z.: An improved nsga-iii algorithm for feature selection used in intrusion detection. Knowl. Based Syst. **116**, 74–85 (2017)
14. Robnik, Ikonja, M., Kononenko, I.: Theoretical and empirical analysis of relieff and rrelieff. Mach. Learn. **53**(1-2), 23–69 (2003)
15. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. **27**(8), 1226–1238 (2005)

16. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley. Tsinghua University Press, (1991)
17. Scholkopf, B., Platt, J., Hofmann, T.: Sparse multinomial logistic regression via bayesian l1 regularisation. In: International Conference on Neural Information Processing Systems. pp. 209–216 (2006)
18. Dong, H., Li, T., Ding, R., Sun, J.: A novel hybrid genetic algorithm with granular information for feature selection and optimization. Appl. Soft. Comput. **65**, 33–46 (2018)
19. Ot, A., Ttn, B., Sm, C.: A novel wrapper-based feature subset selection method using modified binary differential evolution algorithm. Information Sciences **565**, 278–305 (2021)
20. Yang, C.S., Chuang, L.Y., Ke, C.H., Yang, C.H.: Boolean binary particle swarm optimization for feature selection. In: Evol. Comput. pp. 2093–2098 (2008)
21. Tabakhi, S., Moradi, P., Akhlaghian, F.: An unsupervised feature selection algorithm based on ant colony optimization. Eng. Appl. Artif. Intel **32**(6), 112–123 (2014)
22. Kl, F., Kaya, Y., Yildirim, S.: A novel multi population based particle swarm optimization for feature selection. Knowledge Based Systems **219**(4), 1–14 (2021)
23. Zhang, Q., Liu, W., Li, H.: The performance of a new version of moea/d on cec09 unconstrained mop test instances. In: Evolutionary Computation, 2009. CEC '09. IEEE Congress on. pp. 203–208 (2009)
24. Das, A., Das, S.: Feature weighting and selection with a pareto-optimal trade-off between relevancy and redundancy. Pattern Recognit. Lett. **88**, 12–19 (2017)