Applied Intelligence (2020) 50:1272–1288
https://doi.org/10.1007/s10489-019-01597-z

Feature redundancy term variation for mutual information-based feature selection

Wanfu Gao[1,2,3] · Liang Hu[1,2] · Ping Zhang[1,2]

Published online: 10 January 2020

Abstract

Feature selection plays a critical role in many applications that are relevant to machine learning, image processing and gene expression analysis. Traditional feature selection methods intend to maximize feature dependency while minimizing feature redundancy. In previous information-theoretical-based feature selection methods, feature redundancy term is measured by the mutual information between a candidate feature and each already-selected feature or the interaction information among a candidate feature, each already-selected feature and the class. However, the larger values of the traditional feature redundancy term do not indicate the worse a candidate feature because a candidate feature can obtain large redundant information, meanwhile offering large new classification information. To address this issue, we design a new feature redundancy term that considers the relevancy between a candidate feature and the class given each already-selected feature, and a novel feature selection method named min-redundancy and max-dependency (MRMD) is proposed. To verify the effectiveness of our method, MRMD is compared to eight competitive methods on an artificial example and fifteen real-world data sets respectively. The experimental results show that our method achieves the best classification performance with respect to multiple evaluation criteria.

Keywords Machine learning · Feature selection · Information theory · Feature redundancy

1 Introduction

Feature selection is widely used in data preprocessing [13, 20] and intends to select the most informative feature subset from an original feature set [10, 19, 30]. Feature selection methods can reduce the computational cost of data analysis and improve the classification performance [11, 26]. Therefore, feature selection methods have received increasing attention.

According to different selection strategies, feature selection methods can be divided into three models [5, 8, 27]: filter models, wrapper models, and embedded models. Filter models are independent of any classifier and they evaluate a feature based on a specific criterion. Wrapper models are dependent on classifiers and embedded models select the feature subset during the learning stage. Compared to wrapper models and embedded models, filter models have lower computational cost [12, 24]. In this study, we focus on filter models.

There are many techniques that are applied during the feature selection process, such as similarity learning, sparse learning, statistics and information theory [15]. Information-theoretical-based feature selection methods are the focus in the area of data preprocessing [4, 28].

To capture a compact feature subset from an original feature set, traditional feature selection methods focus on minimizing the feature redundancy while maximizing the feature dependency. However, a critical issue is that the larger values of the traditional feature redundancy term do not indicate the worse a candidate feature because a

✉ Ping Zhang
zhangping18@mails.jlu.edu.cn

Wanfu Gao
gaowf@jlu.edu.cn

Liang Hu
543786450@qq.com

1 College of Computer Science and Technology, JiLin University, Changchun, People's Republic of China

2 Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, People's Republic of China

3 College of Chemistry, Jilin University, Changchun, People's Republic of China

candidate feature can obtain large redundant information, meanwhile offering large new classification information.

As described in Fig. 1, X_{k1} and X_{k2} are two candidate features, X_j is an already-selected feature and Y is the class. One of the traditional feature redundancy term is the mutual information between the candidate feature and the already-selected feature, and the feature redundancy term can be represented as the union areas of 1 and 3 and the union areas of 5 and 6 in the Fig. 1. Another one of the traditional feature redundancy term is the interaction information among the candidate feature, the already-selected feature and the class, and the feature redundancy term can be represented as the area 3 and the area 6. Obviously, the area 3 is larger than the area 6 and the union areas of 1 and 3 are larger than the union areas of 5 and 6. That is, the redundancy of feature X_{k1} is larger than the feature X_{k2}. However, we can discover that the area 2 is larger than the area 7, that is, X_{k1} offers more new classification information than that of X_{k2} with respect to the class Y. In fact, the candidate feature X_{k1} is more important than the feature X_{k2}.

To address this issue, we propose a novel feature redundancy term and we select the most informative features by minimizing the new feature redundancy term while maximizing the feature dependency term.

The main contributions in this paper are as follows:

(1) In this paper, we design a new feature redundancy term that considers the relevancy between a candidate feature and the class given each already-selected feature.
(2) We discuss the case that the candidate feature is independent of the already-selected feature, in other words, the candidate feature does not contain any redundant information when the candidate feature is independent of each already-selected feature.
(3) We propose a new feature selection method that maximizes feature dependency while minimizing the new feature redundancy term named min-redundancy and max-dependency (MRMD).
(4) We execute MRMD method and eight compared methods on an artificial example to show the feature selection process of different feature selection methods.
(5) Finally, MRMD method is compared to eight feature selection methods on fifteen real-world benchmark data sets that are from different research areas to verify the effectiveness of our method.

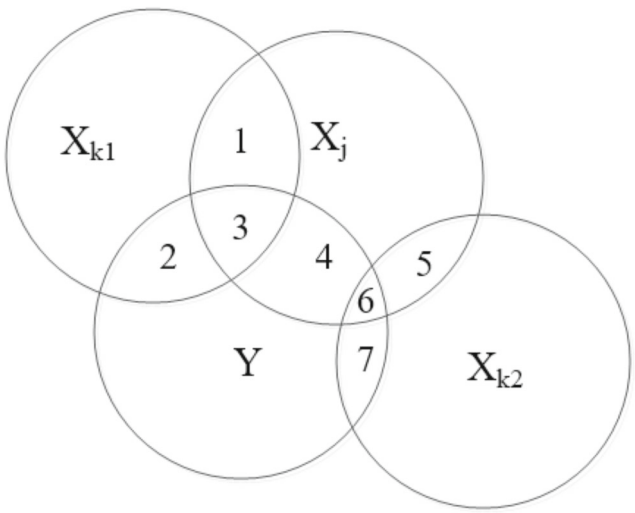

Fig. 1 Venn diagram illustrates the relation between the feature and the class

We organize this paper as follows: Section 2 states theoretical background for this work. In Section 3, we review previous related work. Section 4 proposes a novel feature selection method. In Section 5, we describe and discuss the experimental results on an artificial example and fifteen real-world data sets respectively. Finally, Section 6 concludes this work and gives a plan for future research.

2 Information theory

In this section, we introduce some basic concepts of information theory. There are many metrics in information theory, such as mutual information, conditional mutual information and joint mutual information. Let X, Y and Z be three random variables. Mutual information is defined as follows [7]:

$$I(X;Y) = H(Y) - H(Y|X) \tag{1}$$

where $H(Y)$ is entropy, entropy is used to measure the uncertainty of a random variable, and $H(Y|X)$ represents the conditional entropy that describes the amount of uncertainty left in Y when another random variable X is given. Therefore, mutual information measures the uncertainty reduced when another variable is given.

Conditional mutual information measures the mutual information between two random variables when another variable is introduced, and it is described by Formula (2):

$$I(X;Y|Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z) \tag{2}$$

where $H(X,Y|Z)$ is also a conditional entropy, which measures the amount of uncertainty left in (X,Y) when the variable Z is given.

Another important concept is joint mutual information which is defined as follows:

$$I(X,Y;Z) = I(X;Z) + I(Y;Z|X) \tag{3}$$

Joint mutual information measures the mutual information between (Y,Z) and X.

Interaction information measures the mutual information among three random variables, and it is defined as follows:

$$\begin{aligned} I(X;Y;Z) &= I(X;Y) - I(X;Y|Z) \\ &= I(X;Z) + I(Y;Z) - I(X,Y;Z) \end{aligned} \tag{4}$$

3 Related work

In this section, we review the information-theoretical-based feature selection methods. Feature selection intends to select a feature subset from an original feature set, and maximizes the mutual information between feature subset and the class $I(S;Y)$, where S represents the feature subset. However, due to the amount of calculations and the limited number of observations available for the calculation of the high-dimensional probability density function, the accurate calculation of $I(S;Y)$ is impractical [2]. Cumulative summation of already-selected features is used in many feature selection methods [8, 18, 23].

Mutual Information Maximization (MIM) [14] method evaluates the significance of each candidate feature according to the value of mutual information between the candidate feature and the class. The MIM method does not consider the redundancy among features.

Considering feature dependency and feature redundancy, Mutual Information Feature Selection (MIFS) [1] maximizes the feature dependency while minimizing the feature redundancy:

$$J(X_k) = I(X_k;Y) - \beta \sum_{X_j \in S} I(X_j;X_k) \tag{5}$$

where X_k indicates a candidate feature, X_j represents an already-selected feature, and S is the already-selected feature subset. $J(X_k)$ represents a "scoring" criterion that measures how potentially useful a candidate feature X_k may be. The mutual information $I(X_j;X_k)$ evaluates the feature redundancy. The parameter β varies between zero and one, and β is set to be one according to the suggestion of the authors in our experiment.

With the increase of the number of already-selected feature, it becomes more difficult to remove the redundant features than before [2]. Therefore, the parameter β is set to be the inverse of the number of already-selected feature in minimal-redundancy-maximal-relevance criterion (mRMR) [23], the criterion of mRMR is expressed as follows:

$$J(X_k) = I(X_k;Y) - \frac{1}{|S|} \sum_{X_j \in S} I(X_j;X_k) \tag{6}$$

Conditional Infomax Feature Extraction (CIFE) [18] is proposed to obtain the most informative features. The evaluation criterion is defined as follows:

$$J(X_k) = I(X_k;Y) - \sum_{X_j \in S} I(X_j;X_k) + \sum_{X_j \in S} I(X_j;X_k|Y) \tag{7}$$

According to the Formula (4), Formula (7) can be rewritten as follows:

$$\begin{aligned} J(X_k) &= I(X_k;Y) - \sum_{X_j \in S} \{I(X_j;X_k) - I(X_j;X_k|Y)\} \\ &= I(X_k;Y) - \sum_{X_j \in S} I(X_j;X_k;Y) \end{aligned} \tag{8}$$

Different from MIFS method and mRMR method, the interaction information $I(X_j;X_k;Y)$ is considered as the feature redundancy term in the CIFE method.

The criterion of Joint Mutual Information (JMI) [31] employs cumulative summation of joint mutual information to evaluate the significance of the candidate feature. The criterion of JMI is expressed as Formula (9):

$$J(X_k) = \sum_{X_j \in S} I(X_k, X_j; Y) \tag{9}$$

X_j is an already-selected feature. As a result, $\sum_{X_j \in S} I(X_j;Y)$ can be viewed as a constant in the feature selection process. According to the Formula (4), Formula (9) can be rewritten as follows:

$$\begin{aligned} J(X_k) &= \sum_{X_j \in S} \{I(X_k;Y) + I(X_j;Y) - I(X_j;X_k;Y)\} \\ &\propto |S| I(X_k;Y) - \sum_{X_j \in S} I(X_j;X_k;Y) \\ &\propto I(X_k;Y) - \frac{1}{|S|} \sum_{X_j \in S} I(X_j;X_k;Y) \end{aligned} \tag{10}$$

Similar to CIFE method, the JMI method also considers the interaction information as the measurement of the feature redundancy term. The difference is that the JMI method employs the inverse of the number of already-selected feature $\frac{1}{|S|}$ to balance feature relevance term and feature redundancy term.

Gao et al. propose a feature selection method named Dynamic Change of Selected Feature with the class (DCSF) [8] that designs a new term that calculates dynamic change of selected features with the class. The criterion of DCSF is presented as follows:

$$J(X_k) = \sum_{X_j \in S} \{I(X_k;Y|X_j) + I(X_j;Y|X_k) - I(X_j;X_k)\} \tag{11}$$

additionally, Gao et al. propose a new feature selection named Composition of Feature Relevancy (CFR) that

analyses the composition of feature relevancy [9]. The criterion of CFR is defined by:

$$J(X_k) = \sum_{f_j \in S} \{I(X_k; Y|X_j) - I(X_j; X_k; Y)\} \quad (12)$$

Wang et al. [29] propose a new feature selection method named Max-Relevance and Max-Independence (MRI). MRI is defined as follows:

$$J(X_k) = I(X_k; Y) + \sum_{X_j \in S} \{I(X_k; Y|X_j) + I(X_j; Y|X_k)\} \quad (13)$$

MRI is rewritten as follows:

$$\begin{aligned} J(X_k) &= I(X_k; Y) + \sum_{X_j \in S} \{I(X_k; Y|X_j) + I(X_j; Y|X_k)\} \\ &= I(X_k; Y) + \sum_{X_j \in S} \{I(X_k; Y) + I(X_j; Y) - 2I(X_j; X_k; Y)\} \\ &\propto (|S|+1)I(X_k; Y) - 2\sum_{X_j \in S} I(X_j; X_k; Y) \\ &\propto I(X_k; Y) - \frac{2}{|S|+1} \sum_{X_j \in S} I(X_j; X_k; Y) \end{aligned} \quad (14)$$

additionally, Interaction Weight based Feature Selection (IWFS) takes feature interdependency into accounts. First, IWFS [32] defines interaction weight factor as follows:

$$IW(X_k, X_j) = 1 + \frac{I(X_j, X_k; Y) - I(X_k; Y) - I(X_j; Y)}{H(X_k) + H(X_j)} \quad (15)$$

then, the criterion of IWFS is proposed as follows:

$$J(X_k) = \prod_{X_j \in S} IW(X_k, X_j) * (1 + SU(X_k, X_j)) \quad (16)$$

where $SU(X_k, X_j) = \frac{2I(X_k;X_j)}{H(X_k)+H(X_j)}$, it is a normalized measure of mutual information.

We summarize these feature selection methods mentioned above in Table 1. MIM does not consider the redundancy among features. MIFS, mRMR and DCSF employ the mutual information $I(X_j; X_k)$ to measure the feature redundancy. CIFE, JMI, CFR and MRI use the interaction information $I(X_j; X_k; Y)$ to measure the feature redundancy. Since IWFS is a nonlinear method, it does not define explicitly the feature redundancy term. As shown in Table 1, a critical issue is that these methods do not consider the relevancy between a candidate feature and the class given each already-selected feature when the feature redundancy term is calculated. In addition, Appice et al. design an effective method for eliminating redundant Boolean features for multi-class classification tasks, in which a feature is viewed as a redundant feature when it separates the classes worse than another feature or feature set. The critical idea of the method is that: a set of examples with the same class where the number of different features between examples is small. In this paper, we propose a new feature redundancy term where the feature relevancy is considered. Finally, a novel feature selection method named min-redundancy and max-dependency (MRMD) is proposed.

Table 1 The summaries of feature selection methods

Methods	Feature redundancy term
MIM	None
MIFS	$\beta \sum_{X_j \in S} I(X_j; X_k)$
mRMR	$\frac{1}{\|S\|} \sum_{X_j \in S} I(X_j; X_k)$
DCSF	$\sum_{X_j \in S} I(X_j; X_k)$
CIFE	$\sum_{X_j \in S} I(X_j; X_k; Y)$
JMI	$\frac{1}{\|S\|} \sum_{X_j \in S} I(X_j; X_k; Y)$
CFR	$\sum_{X_j \in S} I(X_j; X_k; Y)$
MRI	$\frac{2}{\|S\|+1} \sum_{X_j \in S} I(X_j; X_k; Y)$
IWFS	Inexplicit

4 The proposed method for feature selection

As we know, feature selection aims to select a compact feature subset that has the maximal dependency with respect to the class [3, 25]. Many feature selection methods use information theory to measure the dependency between a candidate feature and the class, they intend to retain the dependent features and eliminate the redundant features.

The feature redundancy term in traditional feature selection methods is measured by the mutual information between a candidate feature and each already-selected feature or the interaction information among a candidate feture, each already-selected feature and the class. However, the larger values of the traditional feature redundancy term do not indicate the worse a candidate feature because a candidate feature can obtain large redundant information, meanwhile offering large new classification information as analysed in Fig. 1. The new classification information is the relevancy between a candidate feature and the class given the already-selected feature subset S. To address this issue, we design a new feature redundancy term that considers the relevancy between a candidate feature and the class under the condition of the already-selected feature subset S. It is defined as follows:

$$I(S; X_k) - I(X_k; Y|S) \quad (17)$$

where $I(S; X_k)$ is traditional feature redundancy term and $I(X_k; Y|S)$ is the new classification information. Suppose that the value of traditional feature redundancy term is large, meanwhile the value of new classification information is large, in such case, $I(S; X_k)$ is overestimated due to the large new classification information. Employing Formula (17) can reduce the negative impact of redundant information $I(S; X_k)$. Because of the impractical calculations

for already-selected feature subset, we replace the already-selected feature subset with each already-selected feature [2]. The specific definition is as follows:

$$\frac{1}{|S|}\sum_{X_j\in S}\{I(X_j;X_k)-I(X_k;Y|X_j)\} \tag{18}$$

$\frac{1}{|S|}$ is the inverse of the number of already-selected features $|S|$, which is used to balance the feature dependency term and the feature redundancy term. The criterion combining the new feature redundancy term with the feature dependency term is named min-redundancy and max-dependency (MRMD):

$$J(X_k)=I(X_k;Y)-\frac{1}{|S|}\sum_{X_j\in S}\{I(X_j;X_k)-I(X_k;Y|X_j)\} \tag{19}$$

According to the Formula (19), $I(X_k;Y)$ is the feature dependency term, and $\frac{1}{|S|}\sum_{X_j\in S}\{I(X_j;X_k)-I(X_k;Y|X_j)\}$ is the feature redundancy term, additionally, we consider the case that the candidate feature is independent of the already-selected feature. According to the information theory [7], it means that $I(X_j;X_k)=0$ and $I(X_k;Y|X_j)=I(X_k;Y)$, and the Formula (19) can be reduced to the following Formula:

$$\begin{aligned} J(X_k) &= I(X_k;Y)-\frac{1}{|S|}\sum_{X_j\in S}\{I(X_j;X_k)-I(X_k;Y|X_j)\} \\ &= 2I(X_k;Y) \\ &\propto I(X_k;Y) \end{aligned} \tag{20}$$

In this case, MRMD is equal to the method MIM [14].

The steps of MRMD are described as follows:

1) (Initialization) Set $F \leftarrow$ "Original feature set", $S \leftarrow$ "empty set".
2) (Calculate the mutual information between the class and each candidate feature) For each feature $X_k \in F$, calculate $I(X_k;Y)$.
3) (Select the first feature) Find the feature X_k that with the maximal $I(X_k;Y)$, $F \leftarrow F\setminus\{X_k\}$; $S\leftarrow\{X_k\}$.
4) (Greedy selection) Repeat until $|S|=K$
   (a) (Calculate the new feature redundancy term) For all pairs of features X_k and X_j that $X_k\in F$ and $X_j\in S$, calculate $I(X_j;X_k)$ and $I(X_k;Y|X_j)$.
   (b) (Select the next feature) Choose the feature X_k that maximizes Formula (19). $F\leftarrow F\setminus\{X_k\}$; $S\leftarrow\{X_k\}$.
5) Output the feature set S including the already-selected features.

First, MRMD chooses the maximum of the mutual information between each candidate feature and the class. Then, the new feature redundancy term is calculated to select the feature that maximizes the Formula (19). The procedure is ended until $|S|=K$.

Complexity analysis: suppose that K is the number of features to be selected, M is the number of instances in the data set, and N is the total number of features. The time complexity of mutual information and conditional mutual information is $O(M)$ because all instances need to be examined for probability estimation. In MRMD method, each iteration involves the calculation of the information terms on all features, that is, the time complexity is $O(MN)$, and the total iteration is K times, therefore the time complexity of MRMD is $O(KMN)$. The calculation process of other feature selection methods (CIFE, MIFS, mRMR, IWFS, MRI, DCSF and CFR) are similar to that of MRMD, the time complexity of these methods is $O(KMN)$. The MIM method does not involve the selected features. It only needs to calculate the mutual information between all features and the class one time in the process of feature selection, and then the K top-ranked features on the values of mutual information are the feature subset, that is, the time complexity of MIM is $O(MN)$.

When the candidate features are independent of the already-selected features, that is, $I(f_k;f_j)=0$. MRMD is equal to the MIM, that is, both the two methods select the same features. The difference is that the MIM does not take into account the selected features, while the MRMD includes the process of calculating the mutual information. The time complexity of MIM is $O(MN)$ and the time complexity of MRMD is $O(KMN)$. Therefore, MRMD is more time consuming than MIM. However, when the candidate features are correlation with the already-selected features with respect to the class, MRMD is more effective than MIM.

5 Experimental results and analysis

In this section, we evaluate the effectiveness of MRMD on an artificial example and fifteen real-world data sets respectively. All the experiments are executed on an Intel Core i7 with a 3.40 GHZ processing speed and 16 GB main memory. The programming language is Python, and compared methods CIFE, MIFS, MIM and mRMR come from the scikit-feature feature selection repository [16]. IWFS, MRI, DCSF and CFR methods are achieved by ourselves. Additionally, all the classification algorithms are from scikit-learn repository [22].

5.1 Experiments on an artificial data set

To verify the effectiveness of our method, an artificial example $D=(O,F,Y)$ is presented in Table 2. O

Table 2 An artificial example

objects	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	Y
o_1	1	0	1	1	0	1	1	0	1
o_2	1	0	0	1	1	1	0	1	0
o_3	1	1	0	0	0	1	0	0	0
o_4	0	1	0	1	1	1	1	0	1
o_5	0	1	1	0	0	1	0	1	0
o_6	0	1	0	1	0	1	1	0	1
o_7	1	0	0	0	1	0	0	1	1
o_8	1	0	0	1	1	0	1	0	0
o_9	1	0	1	1	1	0	1	0	1
o_{10}	1	0	1	1	1	0	1	1	1

represents the instance set, F is the feature set and Y represents the class, where $O = \{o_1, o_2, ..., o_{10}\}$, $F = \{X_1, X_2, ..., X_8\}$. CIFE, MIFS, MIM, mRMR, IWFS, MRI, DCSF, CFR and MRMD are executed on this artificial example. We detail the feature selection process of MRMD. $Jr(.)$ represents the value of current feature redundancy term, i.e., Formula (18), and $J(.)$ is the final result of Formula (19). By MRMD, we have

1. The mutual information between X_i and Y is calculated. $I(X_1; Y) = 0.0058, I(X_2; Y) = 0.02, I(X_3; Y) = 0.0464 I(X_4; Y) = 0.0913, I(X_5; Y) = 0.02, I(X_6; Y) = 0.0464 I(X_7; Y) = 0.2564, I(X_8; Y) = 0.02;$ $J(X_1) = 0.0058, J(X_2) = 0.02, J(X_3) = 0.0464 J(X_4) = 0.0913, J(X_5) = 0.02, J(X_6) = 0.0464 J(X_7) = 0.2564, J(X_8) = 0.02.$

   The maximum value $J(X_i)$ of every X_i is compared, and X_7 is selected. Then, the candidate feature set is $\{X_1, X_2, X_3, X_4, X_5, X_6, X_8\}$.
2. When the number of already-selected features is equal to one, $Jr(X_1) = -0.1087, Jr(X_2) = -0.17, Jr(X_3) = -0.1171 Jr(X_4) = 0.5078, Jr(X_5) = -0.17, Jr(X_6) = -0.3926 Jr(X_8) = 0.1784;$

   $J(X_1) = 0.1145, J(X_2) = 0.19, J(X_3) = 0.1635 J(X_4) = -0.4165, J(X_5) = 0.19, J(X_6) = 0.439 J(X_8) = -0.1584.$

   The maximum value $J(X_i)$ of every X_i is compared, and X_6 is selected. Then, the candidate feature set is $\{X_1, X_2, X_3, X_4, X_5, X_8\}$.
3. When the number of already-selected features is equal to two, $Jr(X_1) = 0.0618, Jr(X_2) = 0.125, Jr(X_3) =$

Table 3 Date sets description

Data sets	Instances	Features	Classes	Types	Research areas
Musk1	476	166	2	continuous	Physical
Breast	569	30	2	continuous	life
Pixraw10P	100	10000	10	continuous	image
WarpAR10P	130	2400	10	continuous	image
Leukemia	72	7070	2	discrete	Biological
ALLAML	72	7129	2	continuous	Biological
Prostate_GE	102	5966	2	continuous	Biological
GLI_85	85	22283	2	continuous	Biological
Arcene	200	10000	2	continuous	life
ORL	400	1024	40	continuous	image
Waveform	5000	21	3	continuous	Physical
PCMAC	1943	3289	2	discrete	text
Lung_cancer	32	56	3	discrete	Biological
Colon	62	2000	2	discrete	Biological
Lymphoma	96	4026	9	discrete	Biological

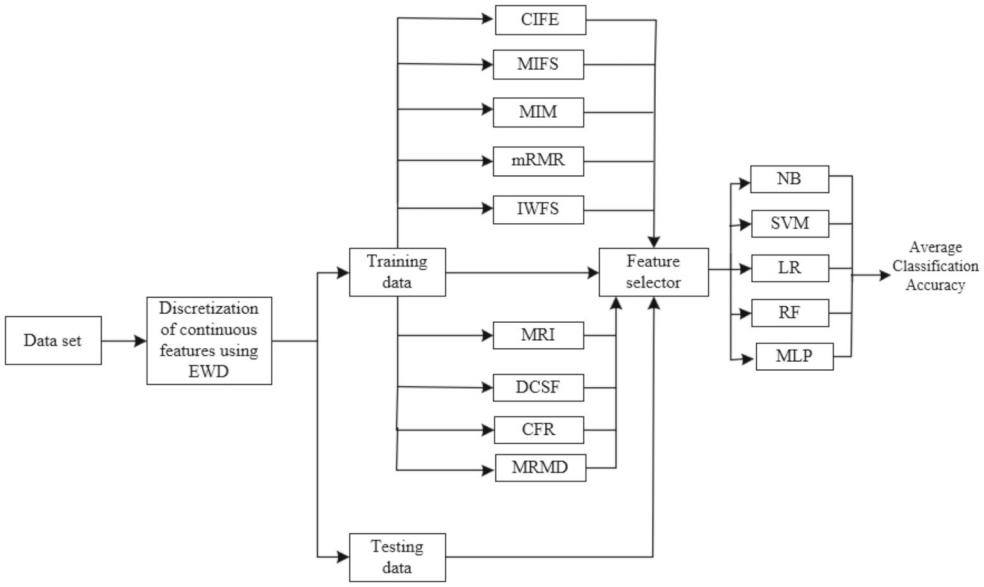

Fig. 2 Experimental framework

$-0.1108 Jr(X_4) = 0.0945, Jr(X_5) = 0.125, Jr(X_8) = -0.1008$;

$J(X_1) = -0.056, J(X_2) = -0.105, J(X_3) = 0.1573 J(X_4) = -0.003, J(X_5) = -0.105, J(X_8) = 0.1208$ The maximum value $J(X_i)$ of every X_i is compared, and X_3 is selected. Then, the candidate feature set is $\{X_1, X_2, X_4, X_5, X_8\}$.

For the artificial example, MRMD selects the feature subset $\{X_7, X_6, X_3\}$, which is the optimal feature subset. The optimal subset can classify all the instances correctly. However, CIFE, IWFS, MRI, DCSF and CFR select $\{X_7, X_6, X_8\}$, mRMR and MIFS select $\{X_7, X_1, X_3\}$, MIM selects select $\{X_7, X_4, X_3\}$. These feature selection methods do not choose the optimal feature subset.

Observing the artificial example above, we discover that the nine feature selection methods choose the same feature X_7 at first. Second, CIFE, MRI, DCSF, CFR and our method choose the feature X_6 and the other two methods, mRMR and MIFS, choose the feature X_1 due to the different feature redundancy terms. Next, MRMD chooses the feature X_3 as the third feature, as a result, MRMD finds the optimal feature subset by only three features. However, CIFE, mRMR, MIM, MIFS, MRI, DCSF and CFR do not choose the optimal feature subset. IWFS does not obtain optimal feature subset when it selects only three features.

5.2 Experimental results on the real-world data sets

In this section, our method is compared to CIFE [18], MIFS [1], MIM [14], mRMR [23], IWFS [32], MRI [29], DCSF [8] and CFR [9] on fifteen real-world data sets that are from different area. Ten-fold cross-validation is used in this experiment. The fifteen data sets come from UCI database [17] and the literature [15] , and the specific descriptions of the data sets are shown in Table 3. These fifteen data sets include binary and multiclass problems, continuous features and discrete features. For the sake of fairness, continuous features are discretized into five bins using equal-width discretization. Additionally, we select these fifteen data sets that come from different research areas, such as physical data, image data and biological data, etc. As a result, the experimental results are convincing. The experimental framework is illustrated in Fig. 2.

We employ five different classifiers, the Naïve-Bayes (NB) classifier, the Support Vector Machine (SVM) classifier, the LogisticRegression classifier (LR), the Random Forest classifier (RF) and the Multi-Layer Perception classifier (MLP), to evaluate the average classification accuracy. The average classification accuracy is obtained across 30 groups of feature subsets on each classifier. The results of the classification performance are recorded in Tables 4, 5, 6, 7 and 8. MRMD implements a paired two-tailed t-test with other compared methods. "+", "-" and "=" indicate that our method performs "better than", "worse than" and "equal to" the corresponding method at a statistical significance level of 5%. The bold fonts indicate the maximal value of the corresponding row. "W/T/L" indicates the number of the data set that our method has higher/equal/lower accuracy than the compared methods.

In Table 4, MRMD method obtains the highest values of the average accuracy on 11 data sets, the average accuracies of MRMD are 66.24%, 92.8%, 91.23%, 57.43%, 95.01%, 93.08%, 90.86%, 85.71%, 70.21%, 55.79%, 76.85%, 73.47%, 52.64%, 83.51% and 64.5% respectively. DCSF achieves the best classification performance on the data set Breast, mRMR obtains the highest values of the average accuracy on the data set ALLAML, MRI outperforms the

Table 4 Average accuracy (mean ± std.) with statistical significance on NB

Data sets	CIFE	MIFS	MIM	mRMR	IWFS	MRI	DCSF	CFR	MRMD
Musk1	58.3±4.91(+)	52.96±4.62(+)	61.33±3.01(+)	57.11±4.27(+)	55.99±1.12(+)	62.17±3.36(+)	63.63±4.01(+)	61.11±3.15(+)	**66.24±1.94**
Breast	90.19±3.25(+)	90.23±2.03(+)	92.26±0.63(+)	92.83±1.03(=)	92.31±0.99(+)	93.09±1.06(-)	**93.31±1.33(-)**	92.66±1.18(=)	92.8±0.99
Pixraw10P	91±8.17(=)	89.67±7.83(+)	78.37±8.92(+)	91.97±8.39(=)	89.47±7.84(+)	88.03±7.53(+)	87.43±9.11(+)	87.47±7.93(+)	**91.23±8.26**
WarpAR10P	44.22±2.9(+)	49.62±4.97(+)	54.72±7.03(+)	57.83±7.59(=)	45.12±4.41(+)	53.38±6.96(+)	51.05±8.77(+)	49.53±7.26(+)	**57.43±7.44**
Leukemia	91.34±1.63(+)	74.24±6.05(+)	94.83±1.5(=)	94.89±1.84(+)	93.86±1.54(+)	94.06±2.03(+)	95.29±1.97(=)	94.77±1.62(=)	**95.01±1.55**
ALLAML	83.09±2.03(+)	81.13±3.1(+)	94.03±3.63(=)	**94.34±4.56(-)**	88.45±1.66(+)	90.4±2.86(+)	83.86±3.11(+)	88.64±5.12(+)	93.08±5.09
Prostate_GE	89.8±2.51(+)	63.93±8(+)	90.65±2.82(=)	88.49±2.96(+)	87.86±2.15(+)	90.26±2.36(=)	89.47±3.31(+)	89.54±2.77(+)	**90.86±2.51**
GLI_85	74.53±1.78(+)	67.49±10.08(+)	86±3.72(=)	83.45±2.15(+)	77.62±1.95(+)	82.45±4.01(+)	78.98±3.1(+)	80.07±1.87(+)	**85.71±3.45**
Arcene	64.76±2.09(+)	55.76±2.57(+)	65.21±0.86(+)	64.98±3.44(+)	63.92±1.81(+)	66.6±1.62(+)	64.24±4.63(+)	67.26±2.26(+)	**70.21±5.4**
ORL	43.92±12.78(+)	55.25±17.19(=)	42.94±11.46(+)	52.24±14.68(+)	42.11±11.7(+)	55.01±17.69(=)	53.58±18.16(+)	51.78±17.68(+)	**55.79±16.46**
Waveform2	75.94±6.3(+)	73.46±7.21(+)	73.55±7.98(+)	76.94±6.54(=)	76.88±6.93(=)	76.85±6.52(=)	76.67±6.58(=)	76.82±6.57(=)	**76.85±6.46**
PCMAC	67.28±2.91(+)	60.49±2.82(+)	73.72±3.14(=)	73.5±3.62(=)	63.84±5.12(+)	**74.48±3.17(-)**	72.98±4.54(=)	74.35±3.4(-)	73.47±3.35
Lung_cancer	**56.25±5.11(-)**	46.67±6.33(+)	51.5±3.83(+)	49.78±5.46(+)	49.89±4.45(+)	55.39±5.5(-)	53.69±6.79(-)	55.47±5.12(-)	52.64±3.58
Colon	72.89±3.71(+)	50.8±11.53(+)	81.48±4.12(+)	81.88±3.01(+)	79.39±3.3(+)	81.53±2.28(+)	82.52±4.59(=)	79.83±2.61(+)	**83.51±3.51**
Lymphoma	45.96±6.68(+)	59.11±9.45(+)	61.69±10.08(+)	65.74±12.21(=)	57.16±10.67(+)	60.82±12.22(+)	58.6±11.46(+)	56.26±11.71(+)	**64.5±13.18**
W/T/L	13/1/1	14/1/0	10/5/0	8/6/1	14/1/0	9/3/3	9/4/2	10/3/2	

Table 5 Average accuracy (mean ± std.) with statistical significance on SVM

Data sets	CIFE	MIFS	MIM	mRMR	IWFS	MRI	DCSF	CFR	MRMD
Musk1	69.35±1.27(=)	**72.29±2.72(-)**	69.22±2.33(+)	70.75±2.76(=)	68.11±3.38(+)	69.71±1.83(=)	71.12±2.6(-)	70.22±1.49(=)	70.25±2.69
Breast	94.1±1.93(+)	94.02±1.69(+)	94.68±1.88(+)	95.8±1.55(=)	94.7±1.72(+)	95.19±1.82(+)	94.99±1.73(+)	94.85±2.1(+)	**95.7±1.52**
Pixraw10P	85.43±7.64(+)	88.43±8.24(+)	75.63±8.56(+)	93.17±9.29(=)	86.5±7.73(+)	**93.53±9.19(-)**	83.57±10.06(+)	89.47±9(+)	92.77±9.3
WarpAR10P	60.13±8.46(+)	71.1±12.29(-)	59.45±10.61(+)	**71.97±12.4(-)**	62.97±12.32(+)	64.63±9.67(+)	57.58±13.4(+)	60.97±11.4(+)	68.98±11.52
Leukemia	88.23±2.89(+)	91.02±1.25(+)	97.26±1.58(=)	96.87±1.53(+)	93.09±1.84(+)	94.73±1.53(+)	94.65±1.66(+)	97±1.9(=)	**97.62±1.83**
ALLAML	80.97±3(+)	86.9±1.39(+)	93.22±2.58(+)	93.63±2.29(+)	82.71±3.66(+)	89.75±3.26(+)	87.88±3.25(+)	89.87±4.35(+)	**94.29±2.86**
Prostate_GE	90.5±2.48(+)	83.05±0.88(+)	91.82±2.95(=)	90.59±3.22(+)	89.1±3.26(+)	89.92±2.41(+)	89.78±3.17(+)	89.78±2.67(+)	**91.7±2.63**
GLI_85	79.29±1.95(+)	76.56±2.65(+)	84.78±3.32(=)	82.74±2.34(+)	79.53±3.26(+)	84.95±3.73(=)	77.87±3.51(+)	80.44±2.53(+)	**84.17±2.4**
Arcene	77.7±5.21(=)	60.31±2.02(+)	71.39±4.31(+)	72.34±6.56(+)	75.39±5.66(+)	76.32±4.79(=)	70.25±5.97(+)	78.45±4.84(=)	**77.17±7.37**
ORL	58.16±15.57(+)	71.73±20.36(+)	65.48±18.75(+)	74.03±21.12(=)	58.54±14.75(+)	71.12±21.35(+)	70.24±22.13(+)	69.28±21.33(+)	**74.09±21.35**
Waveform2	76.44±6.74(+)	74.07±7.68(+)	76.53±8.9(+)	78.67±7.54(+)	77.89±7.05(+)	78.79±7.43(+)	77.56±7.09(+)	78.67±7.54(+)	**78.87±7.42**
PCMAC	78.12±1.74(+)	77.97±1.68(+)	82.91±3.56(+)	**83.96±3.76(-)**	80.03±2.16(+)	82.72±3.3(+)	83.66±4.49(=)	82.52±3.46(+)	83.6±3.56
Lung_cancer	51.28±10.55(+)	40.14±7.15(+)	57.42±6.78(+)	57.69±5.83(+)	50.53±12.59(+)	51.69±9.19(+)	51.11±9.19(+)	51.22±9.17(+)	**59.25±6.35**
Colon	73.38±4.16(+)	74.46±4.68(+)	76.4±2.78(=)	77.14±2.23(=)	71.5±3.85(+)	77.05±1.6(=)	75.23±4.42(+)	75.62±2.7(+)	**76.4±2.98**
Lymphoma	56.05±3.84(+)	61.16±6.7(+)	72.52±8.57(+)	**76.16±8.58(-)**	59.13±7.21(+)	71.29±7.4(+)	72.01±7.17(+)	74.06±9.26(=)	73.84±8.37
W/T/L	13/2/0	13/0/2	11/4/0	7/5/3	15/0/0	10/4/1	13/1/1	11/4/0	

Table 6 Average accuracy (mean ± std.) with statistical significance on LR

Data sets	CIFE	MIFS	MIM	mRMR	IWFS	MRI	DCSF	CFR	MRMD
Musk1	62.66±2.8(+)	70.3±3.62(=)	64.57±4.07(+)	**70.24±3.46(-)**	62.17±1.66(+)	62.05±3.28(+)	66.06±3.03(+)	63.58±3.16(+)	69.43±3.07
Breast	94.14±1.78(+)	94.26±1.94(+)	94.68±1.82(+)	95.93±1.72(=)	94.66±1.66(+)	95.04±1.7(+)	95.04±1.76(+)	94.67±1.98(+)	**95.85±1.61**
Pixraw10P	80.57±6.81(+)	88.23±9.1(=)	66.93±5.39(+)	89.43±8.81(=)	82.93±6.67(+)	**90.2±8.86(-)**	79.17±10.3(+)	84.47±8.45(+)	88.83±8.25
WarpAR10P	57.82±11.22(+)	66.4±14.09(+)	55.17±11.47(+)	**71.75±14.81(-)**	58.02±14.14(+)	61.95±11.32(+)	57.7±13.86(+)	57.3±12.69(+)	69.7±14.4
Leukemia	92.41±2.3(+)	90.68±1.15(+)	95.67±1.86(=)	94.86±1.89(+)	92.37±2.24(+)	95.53±1.72(=)	94.05±2.45(+)	95.59±3(=)	**95.65±2.27**
ALLAML	83.61±2.22(+)	91.03±2.83(+)	92.95±3.47(+)	92.68±3.7(+)	85.79±1.69(+)	90.51±4.46(+)	85.65±3.51(+)	88.88±4.42(+)	**93.6±3.6**
Prostate_GE	88.89±2.97(+)	84.55±1.19(+)	90.63±2.62(=)	90.02±3.11(+)	90.6±2.81(=)	86.12±2.25(+)	88.95±3.95(+)	86.21±2.23(+)	**90.62±3.08**
GLI_85	73.61±2.02(+)	82.6±1.82(+)	82.38±2.82(+)	80.67±2.8(+)	78.65±2.87(+)	81.59±3.23(+)	75.94±2.69(+)	79.32±2.8(+)	**83.5±2.18**
Arcene	72.55±5.07(-)	58.38±1.35(+)	67.59±3.8(+)	70.42±4.98(=)	71.47±3.91(=)	73.11±5.14(-)	65.32±4.98(+)	**74.61±5.67(-)**	70.58±5.65
ORL	48.1±12.78(+)	65.31±18.82(=)	58.1±16.54(+)	65.67±18.75(=)	47.82±12.31(+)	63.23±20.03(+)	61.59±20.68(+)	61.35±20.42(+)	**65.23±18.3**
Waveform2	77.09±6.92(+)	74.5±7.95(+)	76.9±9.1(+)	79.12±7.68(+)	78.33±7.45(+)	79.21±7.6(+)	78.1±7.26(+)	79.09±7.67(+)	**79.29±7.58**
PCMAC	78.81±1.88(+)	77.91±1.65(+)	82.49±3.39(+)	84.18±3.89(+)	80.01±2.14(+)	82.04±3.07(+)	83.48±4.43(=)	81.77±3.25(+)	**83.13±3.35**
Lung_cancer	54.36±5.02(=)	42.86±5.13(+)	**57.06±6.71(-)**	52.56±6.02(=)	54.06±5.11(=)	52.94±8.74(=)	50.06±8.32(+)	52.58±8.57(=)	54.14±4.37
Colon	72.52±3.83(+)	77.93±2.94(+)	78.05±2.63(+)	80.94±2.04(+)	81.71±2.39(=)	80.08±2.29(+)	74.13±2.54(+)	75.39±3.61(+)	**81.61±2.12**
Lymphoma	56.65±4.59(+)	68.92±5.79(+)	73.89±7.72(+)	80.12±9.39(=)	67.47±5.52(+)	77.25±10.19(+)	78.96±11.27(=)	77.31±11.32(+)	**79.61±10.53**
W/T/L	13/1/1	12/3/0	12/2/1	7/6/2	11/4/0	11/2/2	13/2/0	12/2/1	

Table 7 Average accuracy (mean ± std.) with statistical significance on RF

Data sets	CIFE	MIFS	MIM	mRMR	IWFS	MRI	DCSF	CFR	MRMD
Musk1	70.42±2.55(+)	72.74±2.12(=)	69.84±3.56(+)	72.51±2.73(=)	67.82±4.07(+)	68.67±2.58(+)	71.71±2.78(+)	69.78±2.57(+)	**72.87±3.58**
Breast	92.72±1.35(+)	92.72±1.4(+)	94.08±1.26(+)	94.38±1.25(=)	93.7±1.36(+)	93.57±1.4(+)	93.75±1.44(+)	93.67±1.54(+)	**94.41±1.19**
Pixraw10P	89±8.48(+)	91.2±8.31(+)	76.6±8.49(+)	**94.57±9.28(-)**	89.37±8.04(+)	92.73±8.7(+)	83.57±8.89(+)	87.87±8.52(+)	93.93±9.13
WarpAR10P	54.2±5.63(+)	66.72±9.99(+)	60.82±10.04(+)	**70.83±11.84(-)**	52.77±7.71(+)	63.83±8.8(+)	56.58±10.17(+)	59.32±9.2(+)	67.32±10.95
Leukemia	91.73±3.38(+)	92.03±2.03(+)	97.22±1.48(=)	96.7±1.72(+)	93.24±2.5(+)	96.28±2.92(=)	95.08±2.52(+)	96.63±1.38(=)	**97.13±1.44**
ALLAML	81.85±2.5(+)	91.79±1.9(+)	93.93±3.15(=)	94.71±2.51(=)	84.53±2.39(+)	87.06±4.35(+)	84.63±2.17(+)	87.95±2.77(+)	**94.08±3.07**
Prostate_GE	87.81±2.63(+)	82.22±1.08(+)	89.87±2.76(=)	87.96±3.16(+)	88.52±3.17(=)	89±2.27(=)	87.63±3.13(+)	88.22±2.47(=)	**89.17±3.03**
GLI_85	72.31±3.16(+)	75.73±1.32(+)	82.97±3.29(=)	82.93±3(=)	76.63±2.92(+)	81.05±3.76(+)	76.23±4.33(+)	77.37±3.4(+)	**83.02±3.15**
Arcene	76.16±5.4(-)	58.03±1.49(+)	74.03±5.84(-)	71.58±5.41(+)	73.67±5.38(=)	76.12±4.44(-)	69.24±5.17(+)	**77.07±4.7(-)**	72.8±5.31
ORL	49.34±10.92(+)	59.46±14.53(+)	58.12±14.5(+)	63.37±16.31(+)	49.18±10.51(+)	59.7±16.6(+)	57.74±16.45(+)	57.73±16.29(+)	**64.08±16.9**
Waveform2	73.69±5.6(+)	70.67±6.74(+)	73.6±7.43(+)	75.66±6.18(+)	74.92±5.74(+)	75.82±6.01(=)	74.4±5.75(+)	75.66±6.14(+)	**75.99±6.07**
PCMAC	78.62±1.92(+)	78.45±1.81(+)	83.25±3.66(+)	**84.58±4(-)**	80.48±2.28(+)	82.85±3.38(+)	84.31±4.74(=)	82.64±3.55(+)	84.02±3.7
Lung_cancer	49.08±8.13(=)	41.89±9.11(+)	51.81±8.08(=)	49.75±7.62(=)	49.28±7.89(=)	48.72±11.76(=)	49.28±10.33(=)	48.5±8.75(=)	**52.25±9.06**
Colon	72.6±3.3(+)	77.97±4.13(=)	76.75±4.04(+)	78.98±3.49(=)	71.07±4.5(+)	75.21±3.85(+)	76.05±4.65(+)	74.44±5.34(+)	**78.33±3.48**
Lymphoma	57.57±4.75(+)	68.37±6.09(+)	70.33±8.75(+)	75.83±9.04(=)	65.74±5.92(+)	75.65±10.11(=)	74.42±8.59(=)	73.07±9.16(=)	**74.66±10.48**
W/T/L	13/1/1	13/2/0	9/5/1	5/7/3	12/3/0	9/5/1	12/3/0	10/4/1	

Table 8 Average accuracy (mean ± std.) with statistical significance on MLP

Data sets	CIFE	MIFS	MIM	mRMR	IWFS	MRI	DCSF	CFR	MRMD
Musk1	65.71±2.64(+)	69.43±2(=)	67.74±3.46(+)	69.18±1.82(=)	65.15±2.98(+)	66.06±3.18(+)	67.37±2.14(+)	66.77±3.05(+)	**69.01±2.21**
Breast	91.91±1.22(+)	91.48±1.19(+)	93.67±0.99(=)	93.77±1.28(=)	92.42±1.29(+)	92.76±1.12(+)	92.84±1.27(+)	92.37±1.04(+)	**93.76±1.13**
Pixraw10P	75.57±8.08(+)	84.47±7.26(+)	72.07±6.57(+)	87.57±8.38(=)	78.27±7.44(+)	88.23±8.01(=)	74.63±7.17(+)	80.7±7.19(+)	**87.2±7.62**
WarpAR10P	48±4.9(+)	57.85±7.71(=)	51.5±6.6(+)	**61.05±9.09(-)**	48.95±5.89(+)	53.03±6.29(+)	46.88±7.06(+)	51.55±6.44(+)	57.3±7.76
Leukemia	93.21±1.58(+)	91.05±2.24(+)	95.56±1.22(+)	95.46±1.47(+)	94.95±1.47(+)	95.34±1.46(+)	94±1.52(+)	95.68±1.79(+)	**96.32±1.01**
ALLAML	82.74±2.56(+)	91.26±1.5(+)	92.32±2.3(+)	94.04±2.25(=)	85±2.78(+)	89.71±3.3(+)	87.48±3.96(+)	89.12±3.64(+)	**93.61±3.07**
Prostate_GE	87.19±2.65(+)	78.45±3.76(+)	87.22±2.84(+)	87.13±4.44(+)	87.62±3.09(=)	86.97±3.09(+)	88.23±2.93(=)	87.49±2.25(+)	**88.96±4.18**
GLI_85	73.86±3.38(+)	78.57±2.12(+)	78.49±3.35(+)	79.71±3.02(+)	76.76±3.42(+)	79.17±2.36(+)	74.44±3.58(+)	77.74±4.19(+)	**81.15±2.54**
Arcene	74.86±4.66(-)	57.2±2.57(+)	71.3±4.8(+)	69.8±4.77(+)	74.35±5.1(-)	**75.54±4.51(-)**	67.1±4.62(+)	75.32±4.33(-)	73.13±5.58
ORL	29.96±5.23(+)	39.95±6.96(+)	39.95±6.72(+)	41.84±7.74(=)	30.8±5.45(+)	37.91±6.98(+)	37.69±7.64(+)	36.72±7.08(+)	**41.43±7.55**
Waveform2	75.99±6.22(+)	73.52±7.16(+)	75.93±8.25(+)	77.9±7.09(+)	77.18±6.6(+)	78.1±6.94(=)	76.91±6.57(+)	77.84±7.01(+)	**78.11±6.91**
PCMAC	76.64±1.99(+)	78.45±1.78(+)	83.03±3.51(+)	84.68±4.05(=)	80.47±2.26(+)	82.53±3.11(+)	84.36±4.74(=)	82.32±3.26(+)	**84.09±3.71**
Lung_cancer	54.83±7.21(=)	46.81±8.69(+)	53.42±8.15(=)	49.31±7.34(+)	53.39±7.87(=)	54.64±9.3(=)	54.11±8.08(=)	55.56±8.39(=)	**56±6.37**
Colon	66.29±5.58(+)	**76.95±2.86(-)**	73.36±4.86(+)	74.86±3.49(=)	70.56±3.99(+)	72.94±4.78(+)	70.03±4.89(+)	65.94±4.56(+)	74.8±3.95
Lymphoma	50.47±3.1(+)	63.84±5.08(+)	63.43±7.14(+)	73.79±8.89(=)	62.57±5.42(+)	71.96±9.15(=)	71.64±8.55(=)	71.55±8.77(=)	**72.28±10.55**
W/T/L	13/1/1	12/2/1	13/2/0	6/8/1	12/2/1	10/4/1	11/4/0	12/2/1	

other feature selection methods on the data set PCMAC and CIFE obtains the highest values of the average accuracy on the data set Lung_cancer. Similarly, MRMD method obtains the highest values of the average accuracy on 10 data sets, 10 data sets, 11 data sets, 12 data sets in Tables 5-8 respectively. Additionally, the average accuracy of MRMD method is equivalent to the MIM method in some data sets on different classifiers. For example, the average accuracy of MRMD method is equivalent to the MIM method on five data sets using NB classifier and four data sets using SVM classifier, respectively. In the whole, we can conclude that MRMD outperforms the other compared methods in terms of average classification accuracy.

We show the statistical results of "W/T/L" in the Tables 4-8 using Figs. 3, 4, 5, 6 and 7. The X-axis represents the ratio of "W/T/L" that obtained using the classifiers and the Y-axis represents the data sets. As shown in Figs. 3-7, MRMD achieves better or equal performance than compared methods on most data sets.

Furthermore, we show the highest value of the average accuracies attained by five classifiers on each data set using the features selected by each feature selection method in Table 9. Table 9 indicates that MRMD obtains the highest value of the highest accuracies on 8 data sets. CFR obtains the highest accuracy on the data sets Leukemia, mRMR achieves the highest accuracies on four data sets, MIM obtains the highest accuracy on data set Colon, and MRI achieves the highest accuracy on the data set Lung_cancer. As a result, MRMD method achieves the best classification performance in terms of the highest accuracies among the 9 feature selection methods.

To show the classification performance clearly, Fig. 8 illustrates the classification accuracies against the number of features on the data sets Musk1 and ORL; The number K on horizontal axis indicates the first K features with an already-selected order. The vertical axis is the average accuracies of the five classifiers at the first K features. Classification accuracy does not increase with the number of selected features, and we set the number of selected feature K from 1 to 30, it is also used in other literature [32]. Different colors and shapes represent different feature selection methods. As shown in Fig. 8, MRMD method outperforms the other compared methods significantly.

Furthermore, the Areas Under ROC Curves (AUCs) and F1 score of the five classifiers for the nine methods are shown by the boxes in Fig. 9, 10, 11, 12 and 13. The AUC represents the area under the ROC curve and is an effective metric to evaluate classification performance. F1 score is another metric of a test's accuracy, where the F1 score reaches its best value at 1 and worst at 0.

As shown in Figs. 9-13, MRMD outperforms the other eight methods in terms of AUCs and F1 score. The classification performance of feature selection methods are

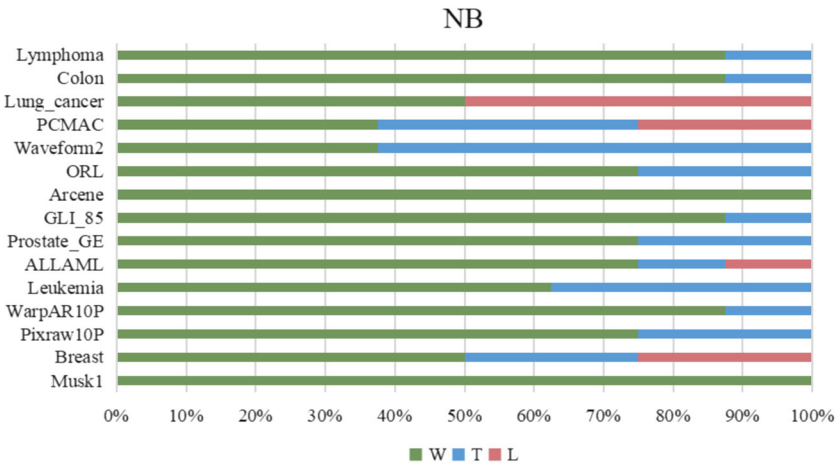

Fig. 3 MRMD performs "better/equally-well/worse" than the compared methods on NB classifier

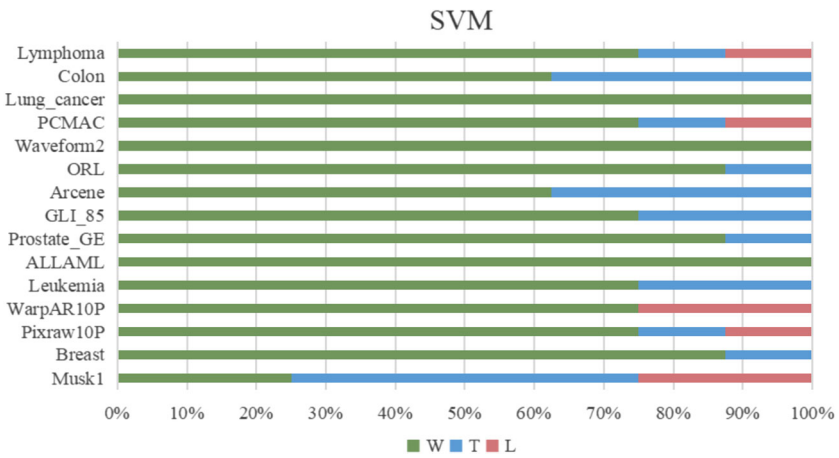

Fig. 4 MRMD performs "better/equally-well/worse" than the compared methods on SVM classifier

Fig. 5 MRMD performs "better/equally-well/worse" than the compared methods on LR classifier

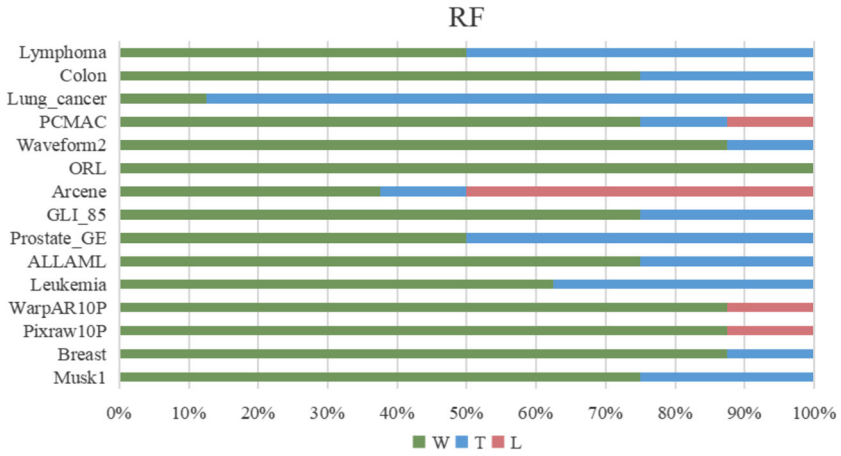

Fig. 6 MRMD performs "better/equally-well/worse" than the compared methods on RF classifier

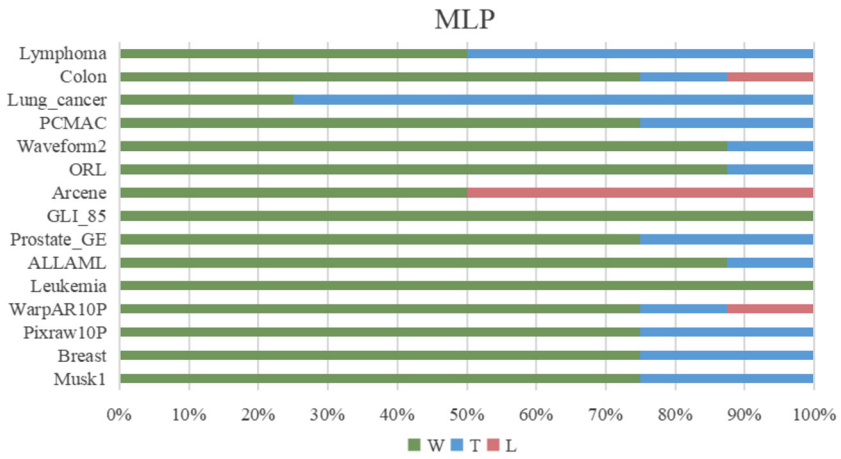

Fig. 7 MRMD performs "better/equally-well/worse" than the compared methods on MLP classifier

Table 9 The highest value of average classification accuracies (%) of the five classifiers on fifteen data sets by each feature selection method

Data sets	CIFE	MIFS	MIM	mRMR	IWFS	MRI	DCSF	CFR	MRMD
Musk1	72.00	72.00	72.00	72.00	66.86	72.00	72.00	72.00	**73.29**
Breast	94.74	94.60	95.16	**95.41**	94.88	94.91	95.09	95.05	95.16
Pixraw10P	92.40	92.00	78.00	**96.20**	90.40	94.40	89.20	90.80	95.40
WarpAR10P	57.90	71.60	65.70	**74.80**	62.40	66.50	66.10	64.00	72.70
Leukemia	93.57	91.92	97.92	97.92	95.92	96.83	97.46	**98.17**	97.67
ALLAML	84.49	90.14	95.21	95.89	87.12	92.93	90.06	92.96	**96.46**
Prostate_GE	90.84	84.62	92.31	92.67	91.67	91.42	90.95	90.25	**93.22**
GLI_85	76.96	78.94	85.33	84.65	80.86	85.35	81.92	82.45	**86.21**
Arcene	77.72	62.24	74.26	75.71	75.28	77.60	73.28	77.64	**77.88**
ORL	55.65	70.30	66.15	70.35	55.40	70.20	72.25	69.55	**72.40**
Waveform2	82.07	82.16	82.31	82.35	82.02	82.32	82.26	82.29	**82.44**
PCMAC	78.26	76.74	83.84	**84.87**	79.62	83.44	84.60	83.48	84.34
Lung_cancer	64.67	57.50	65.17	57.83	65.83	**66.50**	65.33	65.17	61.17
Colon	80.24	81.33	**84.00**	83.00	83.33	80.24	80.24	80.24	81.67
Lymphoma	57.01	69.04	75.31	80.57	68.04	78.99	77.44	78.59	**80.80**

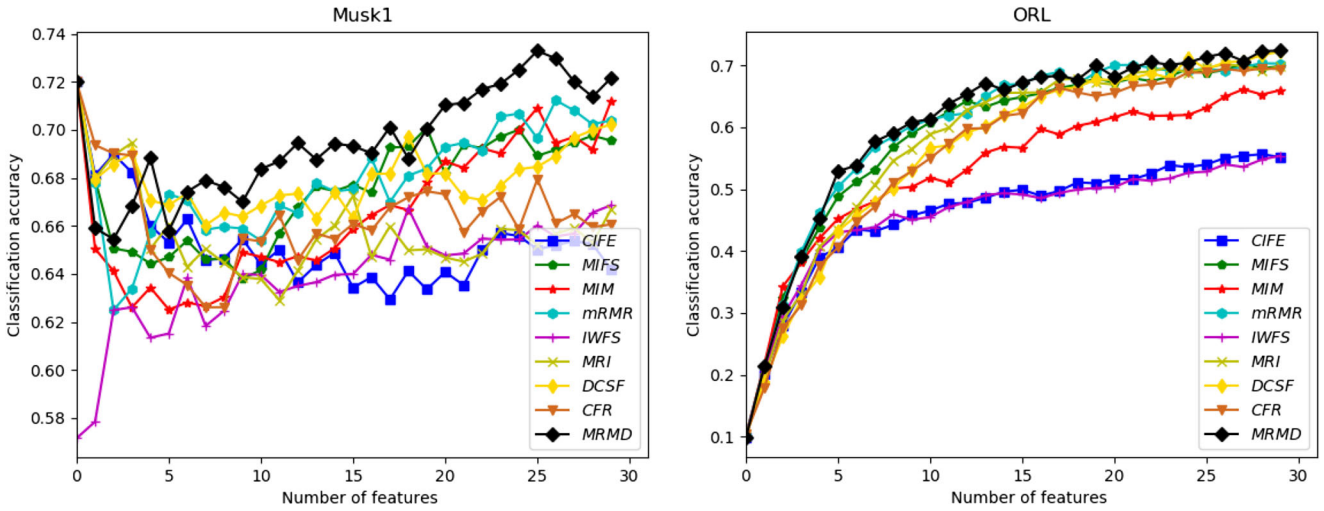

Fig. 8 Average classification accuracy achieved with Musk1 and ORL

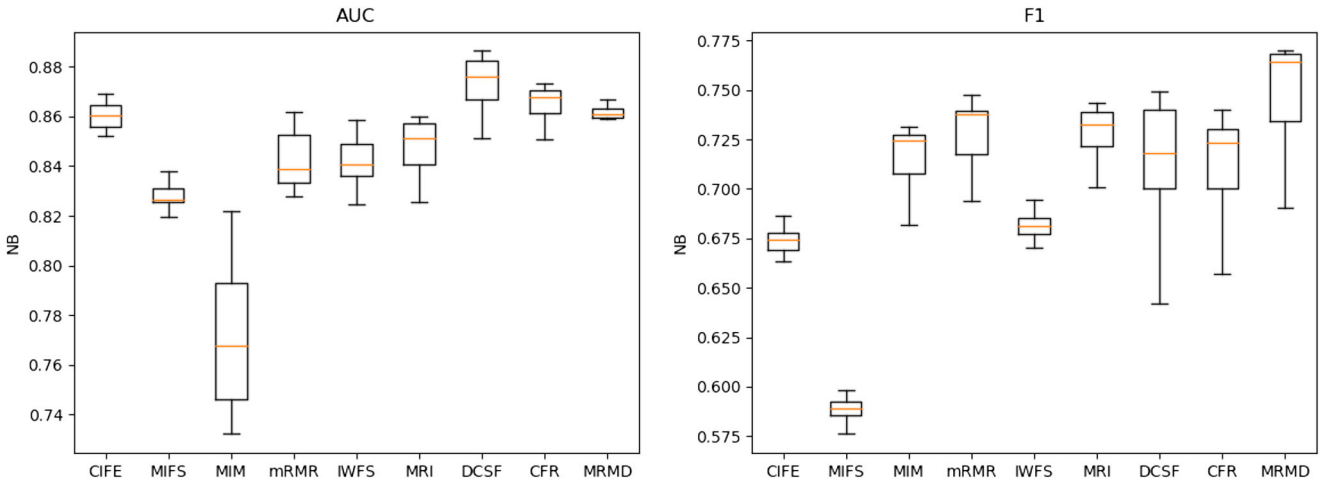

Fig. 9 The AUCs and F1 score across all of the benchmark data sets on NB classifier

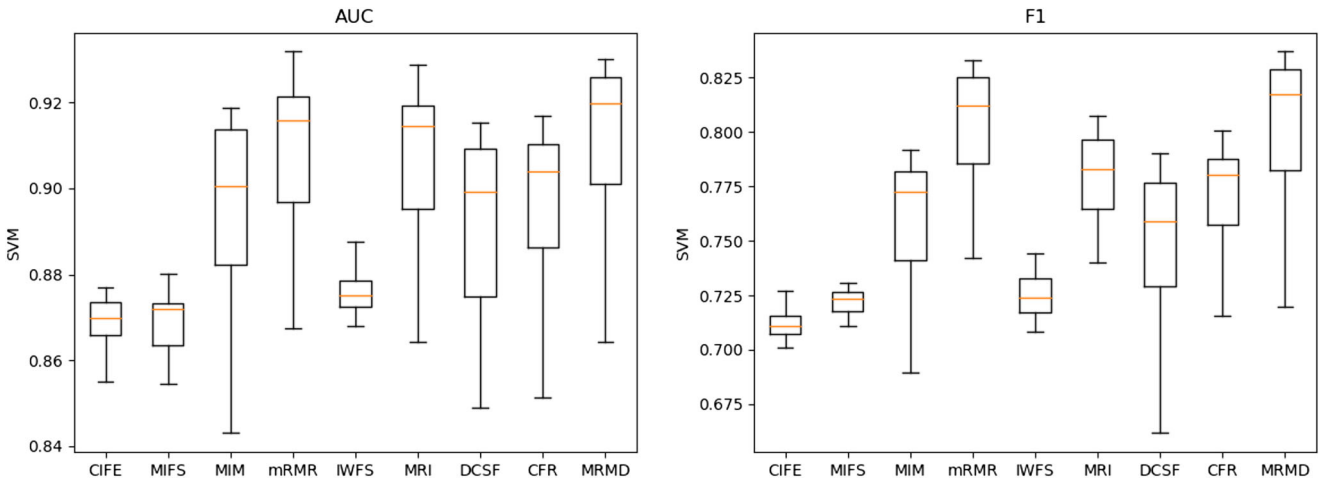

Fig. 10 The AUCs and F1 score across all of the benchmark data sets on SVM classifier

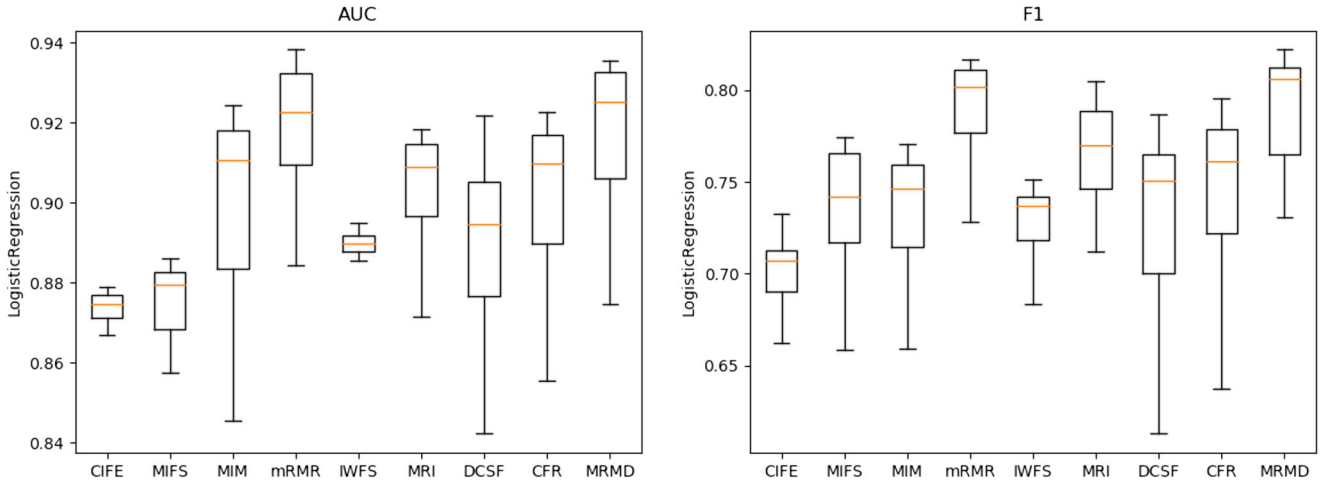

Fig. 11 The AUCs and F1 score across all of the benchmark data sets on LR classifier

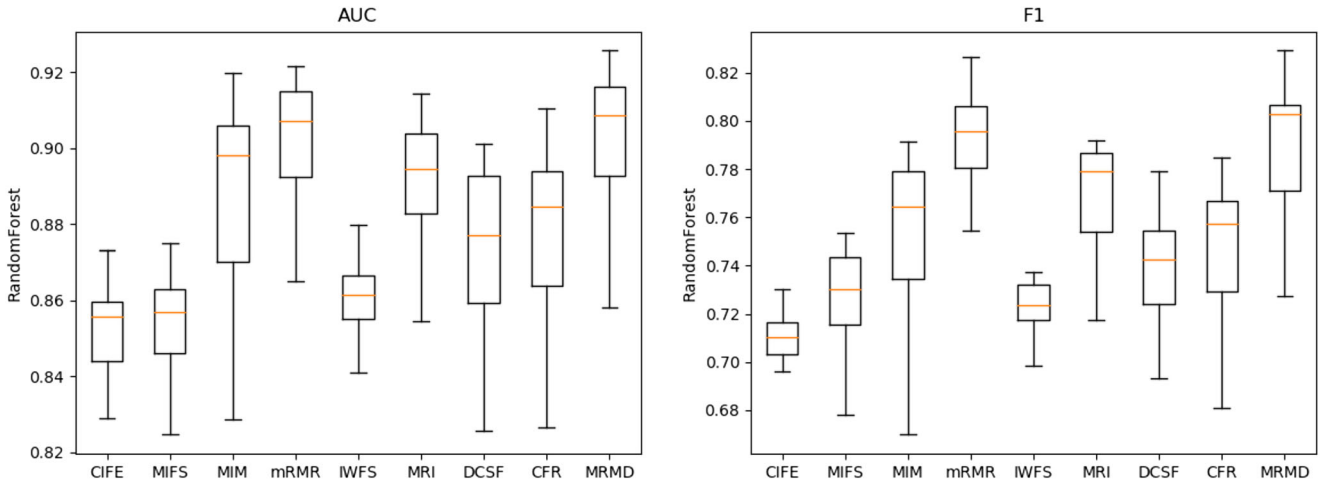

Fig. 12 The AUCs and F1 score across all of the benchmark data sets on RF classifier

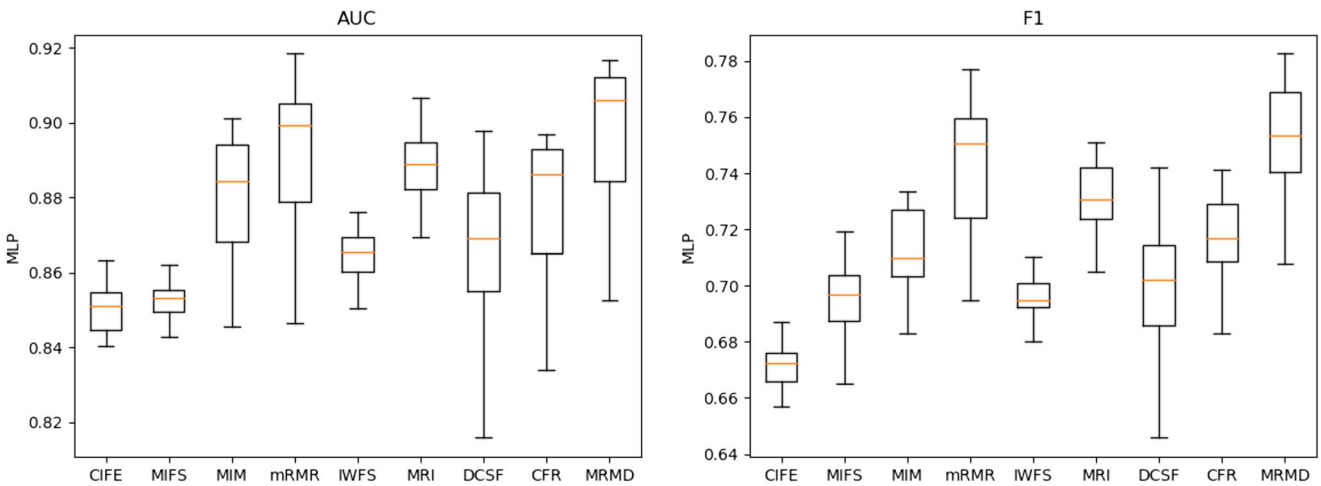

Fig. 13 The AUCs and F1 score across all of the benchmark data sets on MLP classifier

Table 10 The running time (in seconds)

Data sets	CIFE	MIFS	MIM	mRMR	IWFS	MRI	DCSF	CFR	MRMD
Musk1	6.745	6.706	0.083	6.580	12.976	8.489	10.959	4.262	6.720
Breast	0.781	0.832	0.014	0.778	1.589	1.027	1.329	0.537	0.858
Pixraw10P	119.262	115.900	1.476	111.584	230.729	143.752	184.875	73.988	104.980
WarpAR10P	35.880	34.294	0.407	35.157	70.122	45.825	57.696	24.240	33.295
Leukemia	52.716	52.465	0.652	52.971	106.130	66.580	85.502	34.002	52.544
ALLAML	61.169	56.231	0.688	58.529	116.803	77.039	98.985	39.475	60.661
Prostate_GE	67.534	63.334	0.811	65.981	129.598	85.301	110.960	43.129	67.111
GLI_85	217.106	201.207	2.530	212.282	421.744	277.876	355.917	141.086	216.949
Arcene	215.096	179.615	2.294	182.052	365.862	239.130	310.852	119.877	187.384
ORL	43.939	42.681	0.541	42.398	82.630	53.675	66.930	27.786	40.304
Waveform2	3.696	3.497	0.117	3.464	6.894	4.765	5.603	2.276	3.710
PCMAC	530.484	550.341	6.933	546.157	1101.298	696.580	891.207	350.113	549.376
Lung_cancer	0.232	0.219	0.004	0.223	0.430	0.275	0.346	0.155	0.219
Colon	14.116	14.226	0.181	14.115	28.328	17.922	23.010	9.059	14.421
Lymphoma	45.609	44.981	0.526	44.228	90.907	55.659	70.553	28.111	42.810

dependent on the characteristics of classifiers. The same opinion is illustrated in the literature [6, 21]. Additionally, it is worthy to notice that, for some data sets that include missing values, we replace all missing values with the means of the values of features [32].

Finally, we show the running time of MRMD and other eight compared methods (CIFE, MIFS, MIM, mRMR, IWFS, MRI, DCSF and CFR) on fifteen data sets in Table 10. Compared to the methods IWFS, MRI and DCSF, our method is more computationally efficient. The running time of CIFE, MIFS and mRMR is close to our method MRMD. Although MIM has the less running time than other compared methods, the classification performance of MIM is not as good as its running time, which has been demonstrated in Section 5.1. As a result, the time complexity and the running time of MRMD is acceptable.

6 Conclusion and future work

In this work, we propose a new feature redundancy term that not only involves the redundancy between a candidate feature and each already-selected feature, but also considers the relevancy between a candidate feature and the class given each already-selected feature. The proposed method maximizes feature dependency while minimizing the new feature redundancy. To evaluate the effectiveness of our method, the proposed method is compared to eight competitive methods on an artificial example and fifteen real-world data sets respectively. Multiple evaluation criteria including average classification accuracy, the highest accuracy, AUCs and F1 score demonstrate that our method achieves the best classification performance among the nine feature selection methods.

In the future, we plan to explore the correlations between feature dependency and feature redundancy and test the effectiveness of methods on more real-world data sets. Additionally, we will furthermore explore the classification performance of over-dimensioned data sets that are from different areas for feature selection.

Acknowledgments This work is funded by: Postdoctoral Innovative Talents Support Program under Grant No. BX20190137, and National Key R&D Plan of China under Grant No. 2017YFA0604500, National Sci-Tech Support Plan of China under Grant No. 2014BAH02F00, and by National Natural Science Foundation of China under Grant No. 61701190, and by Youth Science Foundation of Jilin Province of China under Grant No. 20160520011JH and 20180520021JH , and by Youth Sci-Tech Innovation Leader and Team Project of Jilin Province of China under Grant No. 20170519017JH, and by Key Technology Innovation Cooperation Project of Government and University for the whole Industry Demonstration under Grant No. SXGJSF2017-4, and Key scientific and technological R&D Plan of Jilin Province of China under Grant No. 20180201103GX, and China Postdoctoral Science Foundation under Grant No. 2018M631873.

References

1. Battiti R (1994) Using mutual information for selecting features in supervised neural net learning. IEEE Trans Neural Netw 5(4):537–550
2. Bennasar M, Hicks Y, Setchi R (2015) Feature selection using joint mutual information maximisation. Expert Syst Appl 42(22):8520–8532
3. Bennasar M, Setchi R, Hicks Y (2013) Feature interaction maximisation. Pattern Recogn Lett 34(14):1630–1635

4. Bolón-Canedo V, Sánchez-Marono N, Alonso-Betanzos A, Benítez JM, Herrera F (2014) A review of microarray datasets and applied feature selection methods. Inf Sci 282:111–135
5. Chen R, Sun N, Chen X, Yang M, Wu Q (2018) Supervised feature selection with a stratified feature weighting method. IEEE Access 6:15,087–15,098
6. Chen S, Ni D, Qin J, Lei B, Wang T, Cheng JZ (2016) Bridging computational features toward multiple semantic features with multi-task regression: a study of ct pulmonary nodules. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 53–60
7. Cover TM, Thomas JA (2012) Elements of information theory. Wiley, New York
8. Gao W, Hu L, Zhang P (2018) Class-specific mutual information variation for feature selection. Pattern Recogn 79:328–339
9. Gao W, Hu L, Zhang P, He J (2018) Feature selection considering the composition of feature relevancy. Pattern Recogn Lett 112:70–74
10. Gui J, Sun Z, Ji S, Tao D, Tan T (2017) Feature selection based on structured sparsity: a comprehensive study. IEEE Trans Neural Netw Learn Syst 28(7):1490–1507
11. 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
12. Huda S, Yearwood J, Jelinek HF, Hassan MM, Fortino G, Buckland M (2016) A hybrid feature selection with ensemble classification for imbalanced healthcare data: a case study for brain tumor diagnosis. IEEE Access 4:9145–9154
13. Lee S, Park YT, dAuriol BJ et al (2012) A novel feature selection method based on normalized mutual information. Appl Intell 37(1):100–120
14. Lewis DD (1992) Feature selection and feature extraction for text categorization. In: Proceedings of the workshop on Speech and Natural Language. Association for Computational Linguistics, pp 212–217
15. Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, Liu H (2016) Feature selection: A data perspective. arXiv:1601.07996
16. Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, Liu H (2018) Feature selection: a data perspective. ACM Comput Surv (CSUR) 50(6):94
17. Lichman M (2013) UCI machine learning repository. http://archive.ics.uci.edu/ml
18. Lin D, Tang X (2006) Conditional infomax learning: an integrated framework for feature extraction and fusion. In: European conference on computer vision. Springer, pp 68–82
19. Liu M, Xu C, Luo Y, Xu C, Wen Y, Tao D (2018) Cost-sensitive feature selection by optimizing f-measures. IEEE Trans Image Process 27(3):1323–1335
20. Mafarja M, Aljarah I, Heidari AA, Hammouri AI, Faris H, AlaM AZ, Mirjalili S (2018) Evolutionary population dynamics and grasshopper optimization approaches for feature selection problems. Knowl-Based Syst 145:25–45
21. Obozinski G, Taskar B, Jordan M (2006) Multi-task feature selection. Statistics Department, UC Berkeley, Technical Report 2(2.2)
22. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine learning in Python. J Mach Learn Res 12:2825–2830
23. Peng H, Long F, Ding C (2005) 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
24. Sayed GI, Hassanien AE, Azar AT (2019) Feature selection via a novel chaotic crow search algorithm[J]. Neural Comput Appl 31(1):171–188
25. Senawi A, Wei HL, Billings SA (2017) A new maximum relevance-minimum multicollinearity (mrmmc) method for feature selection and ranking. Pattern Recogn 67:47–61
26. Sheikhpour R, Sarram MA, Gharaghani S, Chahooki MAZ (2017) A survey on semi-supervised feature selection methods. Pattern Recogn 64:141–158
27. Singh D, Singh B (2019) Hybridization of feature selection and feature weighting for high dimensional data[J]. Appl Intell 49(4):1580–1596
28. Vergara JR, Estévez PA (2014) A review of feature selection methods based on mutual information. Neural Comput Appl 24(1):175–186
29. Wang J, Wei JM, Yang Z, Wang SQ (2017) Feature selection by maximizing independent classification information. IEEE Trans Knowl Data Eng 29(4):828–841
30. Wang Y, Feng L, Zhu J (2018) Novel artificial bee colony based feature selection method for filtering redundant information. Appl Intell 48(4):868–885
31. Yang HH, Moody J (2000) Data visualization and feature selection: New algorithms for nongaussian data. In: Advances in neural information processing systems, pp 687–693
32. Zeng Z, Zhang H, Zhang R, Yin C (2015) A novel feature selection method considering feature interaction. Pattern Recogn 48(8):2656–2666

Wanfu Gao received his B.E. and Ph.D. degrees in the College of Computer Science from Jilin University in 2013 and 2019. He is doing post-doctoral research in the College of Chemistry in Jilin University. His research interests include feature selection and information theory.

Liang Hu received his M.S. and Ph.D. degrees in Computer Science from Jilin University in 1993 and 1999. He is currently a professor and doctoral supervisor at the College of Computer Science and Technology, Jilin University, China. His research areas are network security and distributed computing, including the theories, models, and algorithms of PKI/IBE, IDS/IPS, and grid computing. He is a member of the China Computer Federation.

Ping Zhang received her B.E degree from Hebei GEO University in 2015. Now, She is working toward the PhD degree in the College of Computer Science, Jilin University. Her research interests include feature selection and information theory.