Applied Intelligence
https://doi.org/10.1007/s10489-020-01936-5

A feature selection algorithm based on redundancy analysis and interaction weight

Xiangyuan Gu[1] · Jichang Guo[1] · Chongyi Li[1] · Lijun Xiao[1]

Accepted: 11 September 2020

Abstract
The performance of some three-dimensional mutual information-based algorithms can be affected, since only relevance and interaction are considered. Aiming at solving the problem, a feature selection algorithm based on redundancy analysis and interaction weight is proposed in this paper. The proposed algorithm adopts three-way interaction information to measure the interaction among the class label and features, and processes features for interaction weight analysis. Then, it employs symmetric uncertainty to measure the relevance between features and the class label as well as the redundancy between features, and selects the features with greater relevance and interaction as well as smaller redundancy. To validate the performance, the proposed algorithm is compared with several feature selection algorithms. Since relevance, redundancy, and interaction analysis are all presented, the proposed algorithm can obtain better feature selection performance.

Keywords Three-way interaction information · Symmetric uncertainty · Redundancy analysis · Feature selection

1 Introduction

As an important way of dimensionality reduction, feature selection employs metrics to measure the original features and selects some features with better performance from them [1–3]. Feature selection can be applied to many fields, such as text processing [4, 5], steganalysis [6, 7], network anomaly detecting [8], and underwater objects classification [9]. Mutual information is a metric in feature selection and there are some algorithms based on mutual information, such as Mutual Information based Feature Selection (MIFS) [10] and Minimal-Redundancy-Maximal-Relevance (mRMR) [11]. They employ mutual information to measure the relevance and redundancy. Since, interaction is not considered, their performance is affected.

Three-dimensional mutual information is a supplement of mutual information and it includes three-way interaction information. Some algorithms based on three-way interaction information are proposed, such as Dynamic Weighting-based Feature Selection Algorithm (DWFS) [12] and Interaction Weight based Feature Selection Algorithm (IWFS) [13]. They adopt symmetric uncertainty to measure the relevance and exploit three-way interaction information to measure the interaction, and select the features with greater relevance and interaction. Their performance can be influenced due to ignoring redundancy.

To promote the performance, relevance, redundancy and interaction are all considered, a feature selection algorithm based on redundancy analysis, and interaction weight is proposed. The algorithm initializes the weight of features, and employs symmetric uncertainty to measure the relevance and redundancy and selects the feature with greater relevance and smaller redundancy. Then, it exploits three-way interaction information to measure the interaction and updates the interaction weight of features. Following that, it uses symmetric uncertainty to measure the relevance and redundancy, and selects the features with greater relevance and interaction as well as smaller redundancy.

Major contributions of the paper are as follows. (1) Interaction factor is proposed to measure the interaction. (2) Symmetric uncertainty is adopted for redundancy analysis. (3) Relevance, interaction and redundancy analysis

✉ Jichang Guo
jcguo@tju.edu.cn

[1] School of Electrical and Information Engineering, Tianjin University, Tianjin, 300072, China

Published online: 07 November 2020

are combined for feature selection. (4) A method of determination of redundancy coefficient's value is given.

The remaining of the paper is organized as follows. Section 2 analyzes related works. The proposed algorithm is presented in Section 3. Section 4 shows experimental results. Section 5 gives conclusions and future work.

2 Related works

Given two discrete random variables y and z, $p(y)$ and $p(z)$ are the probability of y and z respectively. $p(y,z)$ is the joint probability. Mutual information is employed to measure the information that two variables share and $I(Y;Z)$ can be expressed as:

$$I(Y;Z)=\sum_{y\in Y}\sum_{z\in Z}p(y,z)\log\frac{p(y,z)}{p(y)p(z)} \tag{1}$$

The greater mutual information value is, the more information two variables share. There are some mutual information-based feature selection algorithms, such as MIFS [10], mRMR [11], Normalized Mutual Information Feature Selection (NMIFS) [14] and Conditional Mutual Information (CMI) [15]. Their objective functions are presented below.

$$MIFS=\arg\max_{f_i\in X}\left[I(f_i;c)-\beta\sum_{f_s\in S}I(f_i;f_s)\right] \tag{2}$$

$$mRMR=\arg\max_{f_i\in X}\left[I(f_i;c)-\frac{1}{|S|}\sum_{f_s\in S}I(f_i;f_s)\right] \tag{3}$$

$$NMIFS=\arg\max_{f_i\in X}\left[I(f_i;c)-\frac{1}{|S|}\sum_{f_s\in S}\frac{I(f_i;f_s)}{\min(H(f_i),H(f_s))}\right] \tag{4}$$

$$CMI=\arg\max_{f_i\in X}\left[I(f_i;c)-\frac{H(f_i|c)}{H(f_i)}\sum_{f_s\in S}\frac{I(f_s;c)I(f_i;f_s)}{H(f_s)H(c)}\right] \tag{5}$$

Where X is the candidate feature set, S is the selected feature set, β is a parameter, $|S|$ is the number of selected features, f_i is a candidate feature, f_s is a selected feature, and c is the class label. $H(c)$ is the information entropy. $H(f_i|c)$ is the conditional entropy. These algorithms employ mutual information between features and the class label to measure the relevance and exploit mutual information between features to measure the redundancy, and their feature selection performance can be influenced due to ignoring three-dimensional mutual information.

Three-dimensional mutual information contains conditional mutual information and three-way interaction information. Conditional mutual information $I(W;Y|Z)$ is defined as:

$$I(W;Y|Z)=\sum_{w\in W}\sum_{y\in Y}\sum_{z\in Z}p(w,y,z)\log\frac{p(w,y|z)}{p(w|z)p(y|z)} \tag{6}$$

where $p(w|z)$ and $p(y|z)$ are the conditional probability.

Three-way interaction information is an extension of mutual information and $I(W;Y;Z)$ has the following relationship with $I(W;Z|Y)$ and $I(W;Z)$, as shown in [16].

$$I(W;Y;Z)=I(W;Z|Y)-I(W;Z) \tag{7}$$

where $I(W;Y;Z)$ can be positive, negative or zero.

DWFS and IWFS are two algorithms that are based on three-dimensional mutual information. They employ symmetric uncertainty that is a normalization of mutual information to measure the relevance between features and the class label, and utilize three-way interaction information to measure the interaction among features and the class label. Part including symmetric uncertainty and weight update part are the key of these algorithms. The part including symmetric uncertainty and weight update part of DWFS are given in (8) and (9) respectively, and that of IWFS are shown in (10) and (11) respectively.

$$SU(f_i,c)=2\frac{I(f_i;c)}{H(f_i)+H(c)} \tag{8}$$

$$\begin{aligned}1+CR(f_i,f_s)&=1+2\frac{I(f_i;c|f_s)-I(f_i;c)}{H(f_i)+H(c)}\\&=1+2\frac{I(f_i;f_s;c)}{H(f_i)+H(c)}\end{aligned} \tag{9}$$

$$1+SU(f_i,c)=1+2\frac{I(f_i;c)}{H(f_i)+H(c)} \tag{10}$$

$$IW(f_i,f_s)=1+\frac{I(f_i;f_s;c)}{H(f_i)+H(f_s)} \tag{11}$$

where $I(f_i;c|f_s)$ is the conditional mutual information of f_i and c when f_s is known. $I(f_i;f_s;c)$ is three-way interaction information. DWFS and IWFS adopt symmetric uncertainty and three-way interaction information for relevance and interaction analysis. Since the redundancy between features is not considered, they do not achieve suboptimal results.

Furthermore, there are some other feature selection algorithms based on three-dimensional mutual information, such as Joint Mutual Information Maximization (JMIM) [17], Max-Relevance and Max-Independence (MRI) [18], Composition of Feature Relevancy (CFR) [19] and Dynamic Change of Selected Feature with the class (DCSF) [20]. Their objective functions are presented below.

$$JMIM = \arg\max_{f_i \in X} \left[\min_{f_s \in S} \left(I\left(f_i, f_s; c\right) \right) \right] \tag{12}$$

$$MRI = \arg\max_{f_i \in X} \left[I\left(f_i; c\right) + \sum_{f_s \in S} I\left(f_i; c | f_s\right) + \sum_{f_s \in S} I\left(f_s; c | f_i\right) \right] \tag{13}$$

$$CFR = \arg\max_{f_i \in X} \left[\sum_{f_s \in S} I\left(f_i; c | f_s\right) + \sum_{f_s \in S} I\left(f_i; f_s; c\right) \right] \tag{14}$$

$$DCSF = \arg\max_{f_i \in X} \left[\sum_{f_s \in S} I\left(f_i; c | f_s\right) + \sum_{f_s \in S} I\left(f_s; c | f_i\right) - \sum_{f_s \in S} I\left(f_i; f_s\right) \right] \tag{15}$$

where $I(f_i, f_s; c)$ is the joint mutual information of f_i, f_s and c. $I(f_s; c|f_i)$ is the conditional mutual information of f_s and c when f_i is known. These algorithms only consider three-dimensional mutual information among features and the class label. Since, they do not consider three-dimensional mutual information among features, their performance can be affected.

3 The proposed algorithm

This section first gives relevance, interaction and redundancy analysis. Then, the diagram and implementation of the proposed feature selection algorithm is presented.

3.1 Relevance, interaction and redundancy analysis

Relevance presents the relationship between one feature and the class label. Since, it contains more shared information with the class label, the feature that is more relevant to the class label is more important. Symmetric uncertainty is adopted to measure the relevance, and $SU(f_i, c)$ between a candidate feature f_i and c is given in (8).

For two candidate features f_i and f_j, if $SU(f_i, c) > SU(f_j, c)$, since more information can be provided, f_i and c are more relevant.

Interaction presents the relationship among two features and the class label. Due to containing more shared information with selected features and the class label, the feature that has greater interaction with the class label and selected features is more important. Considering that the feature with greater three-way interaction information value cannot guarantee achieving better result, interaction factor that is a normalization method of three-way interaction information is given to measure the interaction, and the expression of $If(f_i, f_s, c)$ among f_i, a selected feature f_s and c is expressed as follows:

$$If(f_i, f_s, c) = 2\frac{I(f_i; f_s; c)}{H(f_i) + H(f_s) + H(c)} \tag{16}$$

$If(f_i, f_s, c)$ can be positive, negative or zero. When it is a positive value, the information that f_i and f_s together provide with c is more than they provide individually, and this case suggests that f_i and f_s are complementary in providing the information with c. When $If(f_i, f_s, c)$ is a negative value, the case indicates that they are redundant in providing the information with c. When $If(f_i, f_s, c)$ is zero, the case suggests that they are independent in providing the information with c.

For features f_i and f_j, if $If(f_i, f_s, c) > If(f_j, f_s, c)$, owing to providing more information, f_i, f_s and c have greater interaction.

Redundancy presents the relationship between features. On account of containing more shared information with selected features, the feature that has greater redundancy with selected features is less important. Symmetric uncertainty is employed to measure the redundancy.

For features f_i and f_j, if $SU(f_i, f_s) > SU(f_j, f_s)$, due to providing less information, f_i and f_s have greater redundancy.

3.2 The proposed algorithm

For having the priority to select the features that have greater relevance with the class label and greater interaction with the class label and selected features as well as smaller redundancy with selected features, the part including symmetric uncertainty and weight update part of the proposed algorithm are presented in (17) and (18) respectively.

$$SU(f_i, c)\left[1 - \alpha SU(f_i, f_s)\right] = \frac{2I(f_i; c)}{H(f_i) + H(c)}\left[1 - \frac{2\alpha I(f_i; f_s)}{H(f_i) + H(f_s)}\right] \tag{17}$$

$$1 + If(f_i, f_s, c) = 1 + \frac{2I(f_i; f_s; c)}{H(f_i) + H(f_s) + H(c)} \tag{18}$$

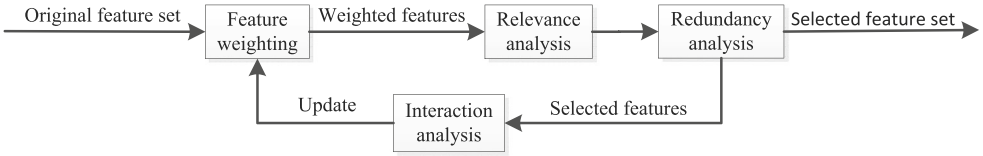


Fig. 1 Diagram of the proposed algorithm

where α is a redundancy coefficient and the value is relevant to the number of dataset's features, $\alpha \in [0, 1]$. In the case that the number of selected features is certain, when the number of dataset's features is a greater value, since there are more candidate features, there exist more symmetric uncertainty values between candidate features and selected features. To guarantee selecting informative candidate features, the difference between values of the second part of (17) should decrease and α is set to a smaller value. Otherwise, the difference between values of the second part of (17) should increase and α is set to a greater value.

In (17), the more features that have greater symmetric uncertainty value with the class label and smaller symmetric uncertainty value with selected features, the better performance algorithm can obtain. In (18), the more features that have greater three-way interaction information value with the class label and selected features, the better result algorithm can achieve.

Based on the analysis above, the diagram of the proposed algorithm is presented in Fig. 1.

As shown in Fig. 1, an original feature set is given and weights of all the features are first initialized. The first part and second part of (17) are adopted to conduct relevance and redundancy analysis for these features respectively and the feature with the maximum of (17) is selected. Then, (18) is employed for interaction analysis and weights of these features are updated. Following that, the first part and second part of (17) are exploited to conduct relevance and redundancy analysis for these updated features separately and the feature with the maximum of (19) is selected. Following the above steps, it selects features until a specific number of features is selected.

$$J(f_i) = \frac{2I(f_i; c)}{H(f_i) + H(c)}\left[1 - \frac{2\alpha I(f_i; f_s)}{H(f_i) + H(f_s)}\right] \times \left[1 + \frac{2I(f_i; f_s; c)}{H(f_i) + H(f_s) + H(c)}\right] \tag{19}$$

3.3 Algorithmic implementation

A feature selection algorithm based on redundancy analysis and interaction weight is proposed and its pseudo-code is shown in Algorithm 1.

Algorithm 1 RAIW: redundancy analysis and interaction weight-based feature selection algorithm.

Input: M: the number of dataset's features, N: the number of features to be selected.
Output: S the selected features.

1: initializes $S = \emptyset$, $X = \{f_1, f_2, \ldots, f_M\}$ and the weight $\omega(f_i)$ for each feature in X to 1;
2: **for** $f_i \in X$ **do**
3: compute $SU(c, f_i)$;
4: **end for**
5: find $f_j \in X$ by maximizing $SU(c, f_i)$;
6: $S = S \cup \{f_j\}$;
7: $X = X - \{f_j\}$;
8: **while** $|S| \leq N$ **do**
9: **for** $f_i \in X$ **do**
10: **for** $f_s \in S$ **do**
11: compute $SU(c, f_i)[1 - \alpha SU(f_i, f_s)]$;
12: compute $If(f_i, f_s, c)$;
13: **end for**
14: update $\omega(f_i)$ by $\omega(f_i) = \omega(f_i)[1 + If(f_i, f_s, c)]$;
15: calculate $J(f_i) = SU(c, f_i)[1 - \alpha SU(f_i, f_s)]\omega(f_i)$;
16: **end for**
17: select the feature $f_k \in X$ that maximizes $J(f_i)$;
18: $S = S \cup \{f_k\}$;
19: $X = X - \{f_k\}$;
20: **end while**

RAIW consists of two parts: in the first part (lines 1-7), S, X and the weight of each candidate feature $\omega(f_i)$ are initialized. Then, symmetric uncertainty between each feature in X and c is calculated, and the feature with the maximum is put into S; in the second part (lines 8-20), testing whether the number of selected features is no less than the predefined number N. If it is less than N, $SU(f_i, f_s)$ and $If(f_i, f_s, c)$ are calculated, and $\omega(f_i)$ is updated. $J(f_i)$ is calculated and the feature with the maximum is put into S. Otherwise, the process ends.

4 Experimental results

To validate the performance of RAIW, DWFS [12], IWFS [13], mRMR [11], JMIM [17], MRI [18], CFR [19] and DCSF [20] are exploited for performance comparisons.

4.1 The datasets and experimental settings

These datasets in Table 1 are from University of California Irvine (UCI) machine learning repository [21] and Arizona State University (ASU) datasets [22]. N is set to 50. Minimum description length method [23] is only utilized for feature selection and exploited to transform these numerical features. Three classifiers J48, IB1 and Naive Bayes are adopted, and their parameters are set to the default values of Waikato Environment for Knowledge Analysis (WEKA) [24]. ASU feature selection software package [25] is adopted.

4.2 Experimental results and analysis

4.2.1 Determination of the α value

Since $\alpha \in [0, 1]$, the α value is set to 0, 0.2, 0.4, 0.6, 0.8 and 1, respectively. The average performance of RAIW with three classifiers is presented in Table 2.

In Table 2, with the increase of dataset's features, the α value that RAIW can obtain better classification accuracy decreases. In these datasets with less features, such as Movement_libras, Musk, Mfeat_fac, Mfeat_pix, Semeion, USPS, madelon and Isolet, RAIW performs the best in the case that α value is 1 except for Mfeat_fac. In the datasets whose features are greater, such as leukemia, Carcinom and arcene, RAIW can achieve better results when α value is 0. In other cases, the α value that RAIW obtains better feature selection performance is uncertainty. For the convenience, the α value is set to 1 in the datasets that the number of features is less than 617, and that is set to 0 in the datasets that the number of features is more than 5748 and that is set to 0.6 in other datasets.

4.2.2 Comparison with other feature selection algorithms

To reduce the impact on randomness, we exploit ten times of 10-fold cross-validation, and take the mean value and standard deviation as the final results [26]. Margins between the performance of RAIW and other algorithms when using three classifiers are shown in Tables 3, 4 and 5. For determining whether the effects of experimental results are obvious, we carry out a one-sided paired t-test at 5% significance level. The number of the datasets that RAIW is superior/equal/inferior to other algorithms is presented by Win/Tie/Loss (W/T/L). Furthermore, we adopt the Friedman's test and the difference among these algorithms is considered to be significant if the p value is less than 0.05. For the results with significant differences, the Nemenyi test is performed to compare RAIW with the other algorithms. The best classification accuracy of these

Table 1 Summary of datasets in the experiment

No.	Datasets	Instances	Features	Classes	Types	Source
1	Movement_libras	360	90	15	Continuous	UCI
2	Musk	476	166	2	Continuous	UCI
3	Mfeat_fac	2000	216	10	Continuous	UCI
4	Mfeat_pix	2000	240	10	Discrete	UCI
5	Semeion	1593	256	10	Discrete	UCI
6	USPS	9298	256	10	Continuous	ASU
7	madelon	2600	500	2	Continuous	ASU
8	Isolet	1560	617	26	Continuous	ASU
9	ORL	400	1024	40	Continuous	ASU
10	colon	62	2000	2	Discrete	ASU
11	warpPIE10P	210	2420	10	Continuous	ASU
12	gisette	7000	5000	2	Continuous	ASU
13	TOX_171	171	5748	4	Continuous	ASU
14	leukemia	72	7070	2	Discrete	ASU
15	Carcinom	174	9182	11	Continuous	ASU
16	arcene	200	10000	2	Continuous	ASU

Table 2 Average performance (%) of RAIW with different α values

Datasets	0	0.2	0.4	0.6	0.8	1	Adopted
Movement_libras	63.88	65.72	66.65	67.18	67.17	67.23	67.23
Musk	79.49	80.68	81.05	81.23	81.26	81.42	81.42
Mfeat_fac	88.51	89.08	89.40	89.59	89.59	89.51	89.51
Mfeat_pix	75.44	77.84	79.62	80.25	80.45	80.50	80.50
Semeion	62.46	66.57	68.28	69.65	71.05	71.38	71.38
USPS	80.16	83.22	84.66	85.21	85.28	85.29	85.29
madelon	69.29	69.46	69.62	69.76	69.87	69.91	69.91
Isolet	72.86	74.06	74.07	74.22	73.91	73.22	74.22
ORL	67.57	68.48	68.84	68.95	69.31	68.86	68.95
s colon	78.43	79.06	79.72	79.37	79.55	79.10	79.37
warpPIE10P	85.83	87.29	87.96	88.53	88.79	88.69	88.53
gisette	91.02	91.45	91.52	91.58	91.65	91.68	91.58
TOX_171	71.18	72.04	72.08	72.69	71.79	71.73	72.69
leukemia	95.85	95.70	95.02	94.27	93.56	93.37	95.85
Carcinom	78.73	78.26	77.89	77.68	77.16	76.22	78.73
arcene	78.46	78.35	78.03	77.95	77.86	77.69	78.46
Avg.	77.45	78.58	79.03	79.26	79.27	79.11	79.60

features that are selected by RAIW and other algorithms with three classifiers is given in Tables 6, 7 and 8. Average performance comparison with three classifiers is presented in Fig. 2.

As shown in Table 3, for the Avg. acc and Avg. rank, DWFS, DCSF and RAIW can yield better results. In terms of the W/T/L values, MRI, CFR and RAIW perform better than other algorithms. Compared with DWFS and IWFS, RAIW can achieve better result in the Avg. acc, Avg. rank and W/T/L values.

In Table 4, the Avg. acc, Avg. rank and W/T/L values show that DWFS, DCSF and RAIW can obtain greater classification accuracy. Different from Table 3, the benefits that RAIW is superior to other algorithms increase in the Avg. acc and RAIW performs better than other algorithms except for DWFS increase in the W/T/L values.

In Table 5, DWFS, DCSF and RAIW can achieve better feature selection performance than other five algorithms in Avg. acc, Avg. rank and W/T/L values. Compared with Tables 3 and 4, the benefits that RAIW performs better than other algorithms except for DCSF gain more increase in the Avg. acc. For the W/T/L values, RAIW has more advantages than other algorithms. In Tables 3, 4 and 5, according to the results of the Friedman test, the difference among eight algorithms is significant with all the three classifiers. On the basis of the results of the Nemenyi test, the difference between RAIW and three algorithms including mRMR, IWFS and JMIM is significant with J48, and the difference between RAIW and other algorithms except for DWFS and DCSF is significant with IB1, and the difference between RAIW and other algorithms except for DCSF is significant with Naive Bayes.

In Tables 6, 7 and 8, in terms of the Avg. acc and Avg. rank, DWFS, DCSF and RAIW can obtain better feature selection performance with three classifiers. Table 6 is taken for example, the number of the datasets that RAIW is superior to other algorithms is 5. Except for RAIW, DCSF performs better than other algorithms and the number of the datasets that DCSF achieves better results is 3. We can know that RAIW has more advantages than other algorithms. In the light of the results of the Friedman test, the difference among eight algorithms is significant in the results evaluated by IB1 and Naive Bayes. According to the results of the Nemenyi test, the difference between RAIW and other algorithms including mRMR, IWFS, MRI and CFR is significant with IB1, and the difference between RAIW and other algorithms except for DCSF is significant with Naive Bayes.

In Fig. 2, RAIW can obtain the desired results, while other algorithms do not handle well in some datasets; for example, DWFS does not achieve better results in Musk, Semeion, USPS and colon. IWFS does not perform well in Mfeat_pix, Semeion, USPS, Isolet and colon.

In comparison with RAIW and other algorithms, mRMR, DWFS and IWFS are taken for example. mRMR only considers the relevance and redundancy. DWFS and IWFS present relevance and interaction analysis, while RAIW considers the relevance, redundancy and interaction. Since

Table 3 Margins (mean ± std. dev.) between the performance of RAIW and other algorithms when using J48

Datasets	RAIW	mRMR	DWFS	IWFS	JMIM	MRI	CFR	DCSF
Movement_libras	64.80±1.52(1)	59.92±0.83(L,6)	61.07±1.06(L,3)	62.90±0.96(L,2)	60.70±0.79(L,4)	59.12±1.61(L,7)	59.04±1.49(L,8)	60.43±1.34(L,5)
Musk	81.45±1.00(1)	80.48±0.89(L,2)	79.68±0.82(L,5)	80.37±1.09(L,4)	79.15±0.95(L,6)	78.85±1.03(L,8)	78.94±1.01(L,7)	80.44±1.19(L,3)
Mfeat_fac	86.08±0.47(1)	85.07±0.28(L,7)	85.53±0.31(L,2.5)	84.27±0.33(L,8)	85.29±0.22(L,4)	85.53±0.21(L,2.5)	85.11±0.27(L,6)	85.18±0.21(L,5)
Mfeat_pix	74.64±0.50(1)	73.05±0.37(L,7)	73.25±0.24(L,6)	71.45±0.47(L,8)	73.56±0.49(L,4)	73.45±0.40(L,5)	73.59±0.37(L,3)	74.28±0.39(L,2)
Semeion	69.62±0.36(1)	62.88±0.35(L,7)	64.16±0.32(L,5)	61.45±0.61(L,8)	65.22±0.38(L,3)	63.75±0.22(L,6)	64.19±0.22(L,4)	66.96±0.45(L,2)
USPS	84.59±0.13(1)	81.32±0.09(L,5)	81.82±0.11(L,4)	81.87±0.09(L,3)	81.16±0.07(L,6)	81.02±0.14(L,8)	81.03±0.12(L,7)	84.25±0.10(L,2)
madelon	75.47±0.56(1)	72.77±0.67(L,8)	74.86±0.51(L,6)	75.18±0.58(L,3)	75.15±0.49(L,4)	75.02±0.56(L,5)	75.19±0.60(L,2)	72.97±0.57(L,7)
Isolet	71.15±0.58(1)	66.14±0.53(L,6)	70.54±0.43(L,2)	63.79±0.43(L,8)	64.24±0.42(L,7)	66.94±0.34(L,4)	66.90±0.33(L,5)	68.65±0.29(L,3)
ORL	51.81±1.13(6.5)	52.34±1.97(T,3)	51.81±1.28(T,6.5)	50.47±1.39(L,8)	52.33±1.30(T,4)	52.82±1.73(W,2)	52.89±1.71(W,1)	52.24±1.48(T,5)
colon	76.55±3.86(4)	77.14±3.31(T,3)	74.32±2.48(L,7)	76.32±2.07(T,5)	77.28±3.77(T,2)	77.30±4.73(T,1)	74.71±3.96(T,6)	73.31±3.77(L,8)
warpPIE10P	79.80±2.24(6)	77.78±1.28(L,8)	80.70±1.89(T,2)	82.51±1.50(W,1)	78.90±0.97(T,7)	80.02±0.63(T,4)	79.93±1.01(T,5)	80.38±0.88(T,3)
gisette	92.40±0.08(5)	92.02±0.08(L,7)	92.66±0.10(W,4)	92.05±0.12(L,6)	91.18±0.12(L,8)	92.84±0.08(W,2)	92.82±0.07(W,3)	93.35±0.08(W,1)
TOX_171	62.21±2.04(2)	58.61±2.21(L,7)	60.94±3.08(T,4)	62.09±2.14(T,3)	57.49±3.01(L,8)	59.78±2.02(L,6)	60.18±1.91(L,5)	62.26±2.71(T,1)
leukemia	93.08±0.67(3)	93.09±0.64(T,2)	92.51±1.14(T,6)	92.05±0.73(L,7)	93.36±0.59(T,1)	92.70±0.98(T,5)	93.07±1.10(T,4)	91.50±1.44(L,8)
Carcinom	71.73±1.80(2)	71.78±2.26(T,1)	69.79±1.58(L,3)	64.47±1.82(L,8)	68.84±1.68(L,4)	68.65±2.20(L,5)	68.01±1.97(L,6)	65.53±1.89(L,7)
arcene	79.09±1.95(1)	77.40±1.85(L,4)	77.43±2.11(L,3)	78.23±1.72(T,2)	73.20±1.84(L,8)	74.47±1.96(L,5)	73.75±1.73(L,6)	73.69±1.40(L,7)
Avg. acc	75.90±10.88	73.86±11.66	74.44±11.34	73.72±11.93	73.57±11.63	73.89±11.59	73.71±11.53	74.09±11.37
Avg. rank	2.34	5.19	4.31	5.25	5	4.72	4.88	4.31
W/T/L	–	12/4/0	11/4/1	12/3/1	12/4/0	11/3/2	11/3/2	12/3/1
Friedman test	$\chi^2 = 16.54$	$p = 0.0206$						
Nemenyi test	$CD = 2.6249$							

Table 4 Margins (mean ± std. dev.) between the performance of RAIW and other algorithms when using IB1

Datasets	RAIW	mRMR	DWFS	IWFS	JMIM	MRI	CFR	DCSF
Movement_libras	81.20±0.61(1)	77.01±0.53(L,6)	78.98±0.62(L,4)	80.06±0.34(L,2)	77.11±0.62(L,5)	76.56±0.75(L,8)	76.70±0.78(L,7)	79.08±0.55(L,3)
Musk	82.22±0.52(1)	81.66±0.75(L,4)	80.62±0.86(L,5)	81.89±1.12(T,2)	79.85±0.85(L,8)	79.94±1.10(L,7)	80.00±1.03(L,6)	81.81±0.90(T,3)
Mfeat_fac	92.11±0.12(1)	91.47±0.10(L,6)	91.73±0.20(L,3)	91.31±0.20(L,8)	91.57±0.15(L,4)	91.56±0.17(L,5)	91.43±0.17(L,7)	91.80±0.18(L,2)
Mfeat_pix	81.29±0.16(1)	73.75±0.15(L,7)	76.83±0.25(L,3)	70.13±0.21(L,8)	75.76±0.39(L,6)	75.79±0.21(L,5)	76.34±0.14(L,4)	79.08±0.24(L,2)
Semeion	71.16±0.30(1)	58.95±0.12(L,7)	61.49±0.26(L,4)	58.09±0.56(L,8)	62.54±0.28(L,3)	60.08±0.24(L,6)	60.65±0.19(L,5)	65.53±0.48(L,2)
USPS	87.73±0.09(2)	82.43±0.09(L,4)	83.10±0.20(L,3)	82.20±0.23(L,6.5)	82.20±0.15(L,6.5)	82.17±0.07(L,8)	82.26±0.09(L,5)	87.89±0.07(W,1)
madelon	70.85±0.61(1)	61.15±0.86(L,8)	69.79±0.62(L,6)	70.26±0.70(L,4)	70.63±0.58(L,2)	70.20±0.72(L,5)	70.50±0.70(L,3)	61.59±0.71(L,7)
Isolet	75.24±0.37(1)	70.09±0.36(L,6)	75.11±0.52(T,2)	66.46±0.55(L,8)	67.53±0.39(L,7)	70.84±0.43(L,5)	70.93±0.43(L,4)	74.22±0.43(L,3)
ORL	81.86±0.74(1)	77.17±0.67(L,8)	79.72±1.06(L,3)	80.78±0.96(L,2)	78.25±1.01(L,5)	77.55±0.57(L,7)	77.61±0.61(L,6)	78.87±0.94(L,4)
colon	78.60±2.10(1)	78.22±1.54(T,2)	75.94±2.32(L,4)	70.87±1.97(L,8)	76.69±2.35(L,3)	71.77±3.41(L,7)	72.24±2.17(L,6)	74.87±1.80(L,5)
warpPIE10P	95.26±0.33(1)	92.39±0.30(L,5)	92.87±0.72(L,3)	93.82±0.56(L,2)	91.98±0.40(L,8)	92.13±0.52(L,7)	92.24±0.49(L,6)	92.69±0.49(L,4)
gisette	91.88±0.08(5)	91.26±0.09(L,6)	92.26±0.07(W,4)	91.05±0.15(L,7)	90.20±0.10(L,8)	92.70±0.06(W,2)	92.58±0.05(W,3)	93.13±0.10(W,1)
TOX_171	85.13±1.26(2)	78.14±1.36(L,7)	85.19±1.30(T,1)	82.59±1.78(L,4)	76.68±1.68(L,8)	81.69±1.64(L,6)	82.05±1.28(L,5)	84.04±1.48(L,3)
leukemia	97.03±0.87(1)	96.16±0.43(L,3)	95.56±0.90(L,6)	88.75±1.88(L,8)	96.61±0.78(T,2)	96.13±0.90(L,4)	95.73±0.95(L,5)	94.22±0.80(L,7)
Carcinom	82.45±1.33(1)	81.39±1.02(L,3.5)	82.32±1.07(T,2)	76.88±1.97(L,8)	81.10±1.06(L,6)	81.39±1.16(L,3.5)	81.35±1.08(L,5)	80.96±1.18(L,7)
arcene	81.58±1.28(3)	81.30±1.02(T,4.5)	82.07±1.70(T,1)	81.30±1.55(T,4.5)	80.50±1.00(L,7)	80.53±1.69(L,6)	80.29±1.68(L,8)	81.62±1.49(T,2)
Avg. acc	83.47±7.79	79.53±10.46	81.47±8.97	79.15±9.91	79.95±9.18	80.06±9.63	80.18±9.43	81.34±9.46
Avg. rank	1.5	5.44	3.38	5.63	5.53	5.72	5.31	3.5
W/T/L	–	14/2/0	11/4/1	14/2/0	15/1/0	15/0/1	15/0/1	12/2/2
Friedman test	$\chi^2 = 44.42$	$p = 1.7745e\text{-}07$						
Nemenyi test	$CD = 2.6249$							

Table 5 Margins (mean ± std. dev.) between the performance of RAIW and other algorithms when using Naive Bayes

Datasets	RAIW	mRMR	DWFS	IWFS	JMIM	MRI	CFR	DCSF
Movement_libras	55.68±1.17(1)	50.81±1.30(L,7)	53.93±1.39(L,3)	52.67±0.94(L,4)	48.57±1.51(L,8)	51.92±1.45(L,6)	52.26±1.50(L,5)	54.45±1.06(L,2)
Musk	80.58±0.58(1)	79.64±0.54(L,2)	78.68±0.49(L,6)	77.41±0.88(L,8)	78.17±0.50(L,7)	78.78±0.55(L,4.5)	78.78±0.58(L,4.5)	79.42±0.64(L,3)
Mfeat_fac	90.35±0.23(1)	88.31±0.11(L,4)	88.20±0.19(L,5)	87.10±0.33(L,8)	87.81±0.17(L,7)	88.45±0.24(L,3)	88.10±0.21(L,6)	89.37±0.15(L,2)
Mfeat_pix	85.57±0.13(1)	79.10±0.11(L,4)	78.45±0.18(L,7)	72.51±0.33(L,8)	79.24±0.24(L,3)	79.03±0.15(L,5)	79.00±0.12(L,6)	81.30±0.12(L,2)
Semeion	73.37±0.17(1)	63.47±0.09(L,7)	65.30±0.08(L,3)	54.78±0.52(L,8)	64.57±0.24(L,4)	64.20±0.14(L,6)	64.52±0.14(L,5)	68.26±0.14(L,2)
USPS	83.55±0.07(1)	79.84±0.04(L,3)	75.69±0.16(L,6)	74.35±0.13(L,8)	75.24±0.06(L,7)	76.07±0.14(L,4)	75.96±0.11(L,5)	82.43±0.09(L,2)
madelon	63.41±0.14(1)	62.85±0.20(L,8)	63.26±0.16(L,5)	63.28±0.16(L,3.5)	63.28±0.15(L,3.5)	63.29±0.16(L,2)	63.24±0.17(L,6)	63.06±0.17(L,7)
Isolet	76.27±0.46(1)	66.87±0.56(L,7)	73.95±0.42(L,2)	68.86±0.43(L,6)	64.40±0.66(L,8)	69.96±0.28(L,5)	70.21±0.23(L,4)	73.74±0.32(L,3)
ORL	73.17±0.62(1)	61.40±1.57(L,8)	71.20±0.86(L,2)	70.84±1.26(L,3)	61.76±1.09(L,7)	65.69±1.44(L,6)	66.17±1.44(L,5)	70.19±0.92(L,4)
colon	82.97±1.34(1)	82.70±1.20(T,2)	80.72±1.47(L,5)	74.39±4.37(L,8)	81.66±1.45(L,4)	79.84±2.40(L,6)	78.95±1.72(L,7)	82.32±2.08(T,3)
warpPIE10P	90.52±0.65(1)	83.27±1.15(L,7)	85.22±1.29(L,3)	84.36±0.80(L,5)	81.60±1.12(L,8)	84.42±0.97(L,4)	84.31±0.97(L,6)	85.92±0.83(L,2)
gisette	90.46±0.13(1)	88.26±0.02(L,3)	87.69±0.08(L,4)	86.23±0.24(L,7)	86.01±0.05(L,8)	87.60±0.04(L,5)	87.48±0.03(L,6)	89.46±0.05(L,2)
TOX_171	70.74±1.07(1)	63.73±1.61(L,7)	68.68±1.26(L,3)	65.64±1.63(L,6)	60.41±2.35(L,8)	66.96±1.56(L,5)	67.04±1.74(L,4)	70.28±1.60(T,2)
leukemia	97.44±0.66(1)	96.27±0.30(L,6)	97.03±0.70(L,3)	95.48±1.96(L,8)	96.18±0.39(L,7)	97.15±0.80(T,2)	96.79±0.71(L,4)	96.70±0.58(L,5)
Carcinom	82.02±0.83(1)	80.23±1.33(L,6)	81.58±0.95(T,2)	76.38±1.85(L,8)	80.19±1.28(L,7)	80.41±0.86(L,4)	80.34±0.87(L,5)	80.46±1.09(L,3)
arcene	74.71±1.02(3)	73.50±1.13(L,4)	74.84±0.94(T,2)	73.18±1.49(L,5)	70.27±0.87(L,8)	71.41±1.06(L,7)	71.45±1.07(L,6)	75.03±1.90(T,1)
Avg. acc	79.43±10.86	75.02±12.30	76.53±10.71	73.59±11.41	73.71±12.43	75.32±11.48	75.29±11.27	77.65±10.80
Avg. rank	1.13	5.31	3.81	6.47	6.53	4.66	5.28	2.81
W/T/L	–	15/1/0	14/2/0	16/0/0	16/0/0	15/1/0	16/0/0	13/3/0
Friedman test	$\chi^2 = 64.12$	$p = 2.2631e\text{-}11$						
Nemenyi test	$CD = 2.6249$							

Table 6 Classification accuracy (%) of selected optimal features when using J48

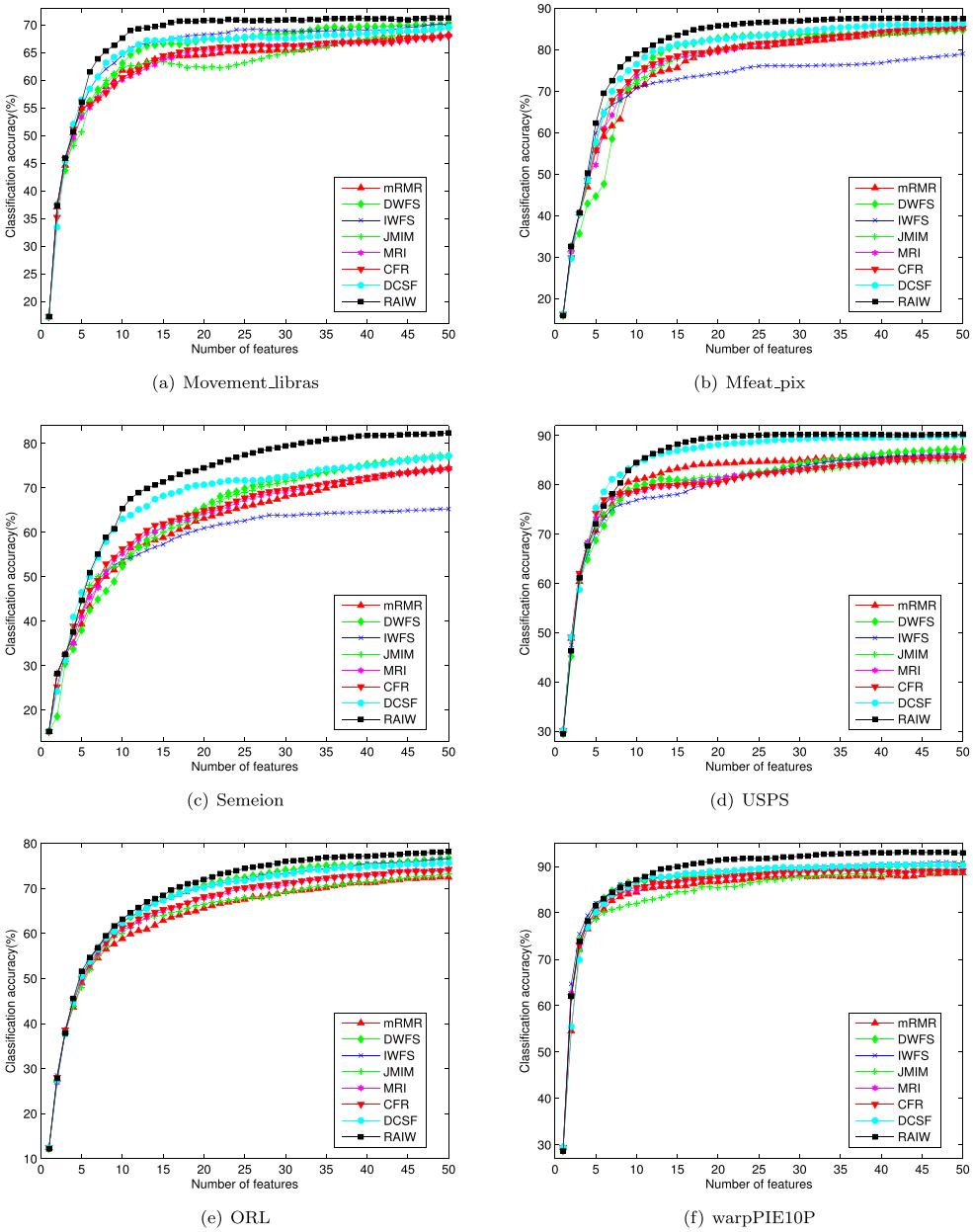
Datasets	RAIW	mRMR	DWFS	IWFS	JMIM	MRI	CFR	DCSF
Movement_libras	69.50 (48,1)	65.61 (50,6)	66.64 (50,4)	67.28 (24,3)	67.36 (49,2)	64.72 (50,7)	64.47 (49,8)	66.25 (50,5)
Musk	83.70 (36,1)	83.05 (49,3)	81.94 (43,6)	82.79 (47,4)	82.33 (33,5)	81.83 (50,7)	81.73 (46,8)	83.16 (35,2)
Mfeat_fac	89.08 (16,2)	88.99 (49,3)	88.43 (35,6)	87.29 (47,8)	89.22 (50,1)	88.68 (34,4)	88.41 (47,7)	88.44 (36,5)
Mfeat_pix	78.35 (43,6)	78.02 (50,7)	78.46 (34,5)	76.87 (50,8)	78.69 (48,4)	79.26 (42,3)	79.29 (42,2)	79.32 (50,1)
Semeion	76.79 (40,1)	71.53 (48,7)	73.07 (49,4)	67.16 (46,8)	74.57 (48,2)	71.84 (50,6)	71.96 (50,5)	73.32 (50,3)
USPS	88.60 (50,1)	84.93 (50,8)	87.26 (50,3)	86.76 (50,4)	85.45 (49,6)	85.46 (50,5)	85.36 (50,7)	88.59 (50,2)
madelon	78.48 (13,6)	75.45 (17,8)	78.51 (12,5)	78.78 (13,1)	78.65 (12,3)	78.53 (13,4)	78.68 (11,2)	75.69 (17,7)
Isolet	78.31 (46,2)	73.41 (47,6)	79.48 (48,1)	70.56 (50,8)	71.76 (50,7)	74.70 (50,5)	74.83 (50,4)	77.95 (50,3)
ORL	54.90 (35,7)	56.70 (47,3)	55.83 (34,6)	53.33 (48,8)	56.02 (47,5)	57.40 (46,1)	57.05 (50,2)	56.08 (45,4)
colon	78.12 (5,3)	78.40 (16,1)	77.86 (4,5)	77.43 (1,6)	77.98 (20,4)	78.36 (7,2)	76.26 (1,7.5)	76.26 (1,7.5)
warpPIE10P	82.43 (13,6)	81.00 (12,8)	83.00 (13,4)	85.14 (42,1)	81.86 (33,7)	83.81 (34,2)	83.43 (30,3)	82.90 (27,5)
gisette	94.19 (50,2)	93.30 (50,6)	93.90 (50,4.5)	93.28 (50,7)	92.93 (47,8)	93.90 (25,4.5)	93.97 (24,3)	94.38 (50,1)
TOX_171	64.75 (38,3)	62.55 (47,7)	64.66 (35,4)	64.95 (34,2)	60.39 (11,8)	63.39 (23,6)	64.57 (25,5)	66.09 (32,1)
leukemia	94.46 (1,4.5)	94.46 (1,4.5)	94.46 (1,4.5)	94.46 (1,4.5)	94.46 (1,4.5)	94.46 (1,4.5)	94.46 (1,4.5)	94.46 (1,4.5)
Carcinom	75.31 (11,2)	76.85 (30,1)	73.75 (9,6)	67.11 (26,8)	74.57 (50,4)	74.75 (35,3)	74.22 (41,5)	73.02 (49,7)
arcene	81.50 (30,1)	80.15 (21,2)	79.40 (16,3.5)	79.40 (18,3.5)	75.50 (7,6)	76.10 (13,5)	75.45 (20,7.5)	75.45 (36,7.5)
Avg. acc	79.28	77.78	78.54	77.04	77.61	77.95	77.76	78.21
Avg. rank	3.03	5.03	4.47	5.25	4.78	4.31	5.03	4.09
Friedman test	$\chi^2 = 10.17$	$p = 0.1791$						

Table 7 Classification accuracy (%) of selected optimal features when using IB1

Datasets	RAIW	mRMR	DWFS	IWFS	JMIM	MRI	CFR	DCSF
Movement_libras	85.56 (17,1)	82.64 (48,8)	85.00 (49,2.5)	85.00 (50,2.5)	83.69 (50,5)	83.11 (50,6.5)	83.11 (50,6.5)	84.50 (49,4)
Musk	85.16 (31,2)	85.04 (45,3)	83.38 (48,5)	84.94 (48,4)	82.96 (50,6)	82.76 (49,8)	82.93 (50,7)	85.54 (50,1)
Mfeat_fac	96.27 (25,3)	96.36 (42,2)	95.95 (49,5)	95.83 (49,6)	96.05 (34,4)	95.74 (27,8)	95.79 (28,7)	96.37 (49,1)
Mfeat_pix	92.80 (50,1)	89.48 (50,7)	90.78 (50,3)	81.34 (50,8)	90.25 (50,6)	90.33 (50,4)	90.29 (50,5)	91.49 (50,2)
Semeion	86.71 (50,1)	77.71 (50,5)	81.43 (50,3)	69.37 (50,8)	81.61 (50,2)	77.01 (50,6)	76.90 (50,7)	81.07 (50,4)
USPS	95.75 (50,1)	89.88 (50,8)	93.33 (50,3)	92.26 (50,4)	90.89 (50,5)	90.47 (50,7)	90.50 (50,6)	95.32 (50,2)
madelon	86.82 (11,4)	67.61 (7,8)	86.58 (12,5)	86.07 (10,6)	86.94 (11,3)	87.03 (11,2)	87.23 (11,1)	67.92 (6,7)
Isolet	84.29 (31,3)	80.16 (50,6)	86.37 (50,1)	74.79 (50,8)	76.85 (50,7)	81.28 (50,5)	81.75 (50,4)	86.33 (50,2)
ORL	92.48 (50,1)	87.23 (50,8)	90.12 (50,3)	91.70 (50,2)	89.70 (50,4)	87.57 (50,7)	87.60 (50,6)	88.92 (50,5)
colon	81.12 (23,1)	80.62 (2,2)	78.60 (2,4)	77.10 (2,7)	78.26 (50,5)	77.57 (2,6)	74.98 (1,8)	78.79 (48,3)
warpPIE10P	99.86 (47,1)	96.14 (50,6)	97.29 (50,3)	98.29 (50,2)	96.33 (48,5)	96.00 (49,7)	95.90 (50,8)	96.62 (48,4)
gisette	95.05 (49,4)	93.62 (36,6)	94.84 (50,5)	93.36 (50,7)	93.08 (48,8)	95.31 (49,3)	95.43 (50,2)	95.92 (48,1)
TOX_171	95.34 (50,2)	86.71 (50,7)	95.84 (50,1)	92.97 (46,4)	85.62 (50,8)	90.97 (50,6)	91.15 (50,5)	93.81 (50,3)
leukemia	99.57 (5,1)	98.36 (10,2)	97.66 (14,6)	96.32 (6,8)	98.34 (29,3)	98.20 (17,4)	97.75 (12,5)	97.11 (50,7)
Carcinom	88.41 (50,6)	88.87 (45,5)	87.95 (38,7)	83.86 (46,8)	89.88 (50,1)	89.16 (47,4)	89.84 (47,2)	89.77 (45,3)
arcene	84.55 (47,3)	84.65 (45,2)	84.25 (45,4)	82.95 (32,6)	83.30 (48,5)	82.35 (26,7)	82.30 (50,8)	84.75 (50,1)
Avg. acc	90.61	86.57	89.34	86.63	87.73	87.80	87.72	88.39
Avg. rank	2.19	5.31	3.78	5.66	4.81	5.66	5.47	3.13
Friedman test	$\chi^2 = 32.38$	$p = 3.4507e-05$						
Nemenyi test	$CD = 2.6249$							

Table 8 Classification accuracy (%) of selected optimal features when using Naive Bayes

Datasets	RAIW	mRMR	DWFS	IWFS	JMIM	MRI	CFR	DCSF
Movement_libras	60.08 (35,1)	56.53 (50,8)	59.39 (39,2)	59.00 (48,3)	56.94 (49,7)	57.19 (28,5)	57.14 (30,6)	58.89 (25,4)
Musk	83.59 (39,1)	82.54 (43,3)	81.64 (50,4)	79.33 (49,8)	80.36 (20,7)	80.61 (30,6)	80.67 (22,5)	82.90 (45,2)
Mfeat_fac	94.61 (48,1)	92.76 (34,4)	91.69 (24,7)	91.14 (46,8)	92.40 (36,6)	92.62 (45,5)	92.80 (48,3)	94.07 (50,2)
Mfeat_pix	91.82 (36,1)	87.89 (50,3)	84.93 (50,7)	78.90 (50,8)	86.87 (50,4)	86.79 (50,5)	86.66 (50,6)	88.18 (50,2)
Semeion	83.78 (50,1)	74.39 (50,5)	76.62 (50,3)	59.42 (49,8)	76.50 (50,4)	74.03 (50,7)	74.05 (50,6)	77.31 (50,2)
USPS	87.29 (19,1)	83.20 (41,3)	81.20 (49,4)	79.84 (50,7)	78.67 (50,8)	81.10 (50,5)	80.95 (50,6)	86.08 (33,2)
madelon	65.15 (4,1)	63.99 (28,6)	64.51 (7,3)	64.05 (3,5)	64.52 (4,2)	64.06 (3,4)	63.71 (4,8)	63.94 (29,7)
Isolet	86.49 (50,1)	74.44 (48,7)	83.42 (50,2)	77.87 (50,6)	70.79 (37,8)	78.69 (50,5)	78.87 (49,4)	83.37 (48,3)
ORL	87.88 (50,1)	74.05 (49,7)	85.20 (50,2)	84.75 (49,3)	73.88 (50,8)	78.05 (50,5)	78.00 (50,6)	82.27 (50,4)
colon	84.57 (40,3)	84.62 (37,2)	81.69 (15,7)	77.90 (3,8)	83.10 (22,4)	81.98 (48,5)	81.83 (37,6)	85.64 (48,1)
warpPIE10P	98.10 (43,1)	90.71 (50,4)	91.33 (50,3)	89.86 (47,7)	89.38 (50,8)	90.48 (50,6)	90.62 (50,5)	92.14 (50,2)
gisette	92.68 (50,1)	88.82 (33,3)	88.37 (9,4.5)	87.26 (15,7)	86.67 (2,8)	88.32 (43,6)	88.37 (44,4.5)	90.80 (50,2)
TOX_171	78.66 (50,1)	67.10 (37,7)	75.60 (50,3)	72.22 (50,6)	64.59 (49,8)	72.37 (43,5)	72.48 (42,4)	77.06 (49,2)
leukemia	98.46 (5,1)	98.07 (10,2)	97.64 (17,6)	96.79 (4,8)	97.50 (6,7)	98.04 (26,4)	98.05 (21,3)	97.80 (12,5)
Carcinom	90.01 (47,2)	88.73 (50,7)	89.20 (44,4)	86.61 (50,8)	88.84 (47,6)	89.45 (50,3)	89.19 (50,5)	90.62 (49,1)
arcene	76.65 (30,2)	74.80 (49,4)	76.30 (32,3)	74.25 (10,5)	71.30 (10,8)	72.30 (50,7)	72.45 (46,6)	77.05 (50,1)
Avg. acc	84.99	80.17	81.80	78.70	78.89	80.38	80.37	83.01
Avg. rank	1.25	4.69	4.03	6.56	6.44	5.19	5.22	2.63
Friedman test	$\chi^2 = 62.26$	$p = 5.3326e-11$						
Nemenyi test	$CD = 2.6249$							



(a) Movement_libras (b) Mfeat_pix

(c) Semeion (d) USPS

(e) ORL (f) warpPIE10P

Fig. 2 Average performance comparisons of algorithms with three classifiers

relevance, redundancy and interaction analysis is presented, RAIW can obtain better feature selection performance than mRMR, DWFS and IWFS.

5 Conclusions and future work

This paper adopts symmetric uncertainty to measure the relevance between features and the class label as well as the redundancy between features, and employs three-way interaction information to measure the interaction among features and the class label. Finally, RAIW is proposed. To verify the effectiveness, RAIW is applied to three classifiers, five UCI datasets and eleven ASU datasets, and the proposed algorithm is compared with other seven feature selection algorithms in terms of their performance. Since relevance, redundancy and interaction are all considered, RAIW performs better than other algorithms.

Although RAIW can achieve better feature selection performance, since only sixteen datasets are exploited in determining redundancy coefficient's value, the redundancy coefficient's value of RAIW is under optimization. To promote feature selection performance, we will further investigate the determination of redundancy coefficient's value in future work.

Acknowledgements This work was supported by the National Natural Science Foundation of China (61771334).

References

1. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
2. Huang XJ, Zhang L, Wang BJ, Li FZ, Zhang Z (2018) Feature clustering based support vector machine recursive feature elimination for gene selection. Appl Intell 48(3):594–607
3. Wang YW, Feng LZ, Zhu JM (2018) Novel artificial bee colony based feature selection method for filtering redundant information. Appl Intell 48(4):868–885
4. Tang B, Kay S, He HB (2016) Toward optimal feature selection in naive bayes for text categorization. IEEE Trans Knowl Data Eng 28(9):2508–2521
5. Shang CX, Li M, Feng SZ, Jiang QS, Fan JP (2013) Feature selection via maximizing global information gain for text classification. Knowl-Based Syst 54:298–309
6. Gu XY, Guo JC (2019) A study on subtractive pixel adjacency matrix features. Multimed Tools Appl 78(14):19681–19695
7. Gu XY, Guo JC, Wei HW, He YH (2020) Spatial-domain steganalytic feature selection based on three-way interaction information and KS test. Soft Comput 24(1):333–340
8. Zhang F, Chan PPK, Biggio B, Yeung DS, Roli F (2016) Adversarial feature selection against evasion attacks. IEEE Trans Cybern 46(3):766–777
9. Fei T, Kraus D, Zoubir AM (2015) Contributions to automatic target recognition systems for underwater mine classification. IEEE Trans Geosci Remote Sens 53(1):505–518
10. Battiti R (1994) Using mutual information for selecting features in supervised neural net learning. IEEE Trans Neural Netw 5(4):537–550
11. Peng HC, Long FH, 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
12. Sun X, Liu YH, Xu MT, Chen HL, Han JW, Wang KH (2013) Feature selection using dynamic weights for classification. Knowl-Based Syst 37:541–549
13. Zeng ZL, Zhang HJ, Zhang R, Yin CX (2015) A novel feature selection method considering feature interaction. Pattern Recogn 48(8):2656–2666
14. Estevez PA, Tesmer M, Perez CA, Zurada JA (2009) Normalized mutual information feature selection. IEEE Trans Neural Netw 20(2):189–201
15. Foithong S, Pinngern O, Attachoo B (2012) Feature subset selection wrapper based on mutual information and rough sets. Expert Syst Appl 39(1):574–584
16. Jakulin A, Bratko I (2004) Testing the significance of attribute interactions. In: Proceedings of international conference on machine learning, pp 409–416
17. Bennasar M, Hicks Y, Setchi R (2015) Feature selection using joint mutual information maximisation. Expert Syst Appl 42(22):8520–8532
18. Wang J, Wei JM, Yang ZL, Wang SQ (2017) Feature selection by maximizing independent classification information. IEEE Trans Knowl Data Eng 29(4):828–841
19. Gao WF, Hu L, Zhang P, He JL (2018) Feature selection considering the composition of feature relevancy. Pattern Recogn Lett 112:70–74
20. Gao WF, Hu L, Zhang P (2018) Class-specific mutual information variation for feature selection. Pattern Recogn 79:328–339
21. Dua D, Graff C (2019) UCI Machine Learning Repository. http://archive.ics.uci.edu/ml
22. Li JD, Cheng KW, Wang SH, Morstatter F, Trevino RP, Tang JL, Liu H (2018) Feature selection: a data perspective. ACM Comput Surv 50(6):1–45
23. Fayyad UM, Irani KB (1993) Multi-interval discretization of continuous-valued attributes for classification learning. In: Proceedings of international joint conference on artificial intelligence, pp 1022–1027
24. Hall MA, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. SIGKDD Explorations 11(1):10–18
25. Zhao Z, Morstatter F, Sharma S, Alelyani S, Anand A, Liu H (2010) ASU feature selection software package. http://featureselection.asu.edu/old/index.php
26. Gu XY, Guo JC, Xiao LJ, Ming T, Li CY (2020) A feature selection algorithm based on equal interval division and minimal-redundancy-maximal-relevance. neural process lett 51(2):1237–1263

Xiangyuan Gu received the Ph.D. degree in the School of Electrical and Information Engeering, Tianjin University, Tianjin, China. His current research focuses on machine learning, steganalysis and especially on feature selection.

Jichang Guo received the M.S. and Ph.D. degrees from the School of Electrical and Information Engeering, Tianjin University, Tianjin, China, in 1993 and 2003, respectively. He is currently a full professor in Tianjin University. His current research interests include digital image processing, video coding and computer vision.

Chongyi Li received the Ph.D. degree from the School of Electrical and Information Engineering, Tianjin University, Tianjin, China, in June 2018. From 2016 to 2017, he was a Joint-Training Ph.D. Student with Australian National University, Australia. He was a Postdoctoral Research Fellow with the Department of Computer Science, City University of Hong Kong (CityU), Hong Kong SAR, China. He is currently a Postdoctoral Research Fellow with the School of Computer Science and Engineering, Nanyang Technological University (NTU), Singapore. His current research focuses on image processing, computer vision, and deep learning, particularly in the domains of image restoration and enhancement.

Lijun Xiao received the M.S. degree in the School of Electrical and Information Engeering, Tianjin University, Tianjin, China. His research interests are computer vision and machine learning.