Applied Intelligence (2022) 52:1436–1447
https://doi.org/10.1007/s10489-021-02412-4

Conditional mutual information-based feature selection algorithm for maximal relevance minimal redundancy

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

Accepted: 3 April 2021 / Published online: 21 May 2021

Abstract

There are many feature selection algorithms based on mutual information and three-dimensional mutual information (TDMI) among features and the class label, since these algorithms do not consider TDMI among features, feature selection performance can be influenced. In view of the problem, this paper investigates feature selection based on TDMI among features. According to the maximal relevance minimal redundancy criterion, joint mutual information among the class label and feature set is adopted to describe relevance, and mutual information between feature set is exploited to describe redundancy. Then, joint mutual information among the class label and feature set as well as mutual information between feature set is decomposed. In the process of decomposing, TDMI among features is considered and an objective function is obtained. Finally, a feature selection algorithm based on conditional mutual information for maximal relevance minimal redundancy (CMI-MRMR) is proposed. To validate the performance, we compare CMI-MRMR with several feature selection algorithms. Experimental results show that CMI-MRMR can achieve better feature selection performance.

Keywords Maximal relevance minimal redundancy · Conditional mutual information · Mutual information · Feature selection

1 Introduction

Feature selection [1–5] aims at selecting some informative features from feature set, which is an important method of dimensionality reduction and has widespread applications, such as text processing [6, 7], steganalysis [8, 9], underwater objects classification [10], network anomaly detection [11], information retrieval [12], and image classification [13, 14]. Feature selection algorithms are divided into three categories: filter, embedded, and wrapper methods. Since classification accuracy of classifier is taken as the metric, embedded and wrapper methods are time-consuming and not robust. Filter methods take less time at the cost of the decline in classification results. It is advisable that the datasets with high dimensional features be dealt with filter methods [15].

The metrics commonly used in filter methods include consistency, distance and MI. Since MI has the capacity of measuring linear and non-linear correlation as well as its invariance under space transformations [16], feature selection based on MI and TDMI is widely investigated. Mutual information maximization (MIM) [17] is a basic feature selection algorithm based on MI. It calculates MI between the class label and features, and selects some features with greater values. On the basis of MIM, feature selection algorithms based on relevance and redundancy are proposed, such as minimum-redundancy maximum-relevance (mRMR) [18], conditional mutual information (CMI) [19], and MIFS-CR [20]. These algorithms adopt MI between features and the class label to describe relevance and exploit MI between features to describe redundancy. Owing to considering relevance and redundancy simultaneously, feature selection performance of these algorithms is improved.

Some feature selection algorithms further consider TDMI to improve the performance and algorithms based on TDMI are proposed, such as interaction weight based feature selection (IWFS) [21], joint mutual information maximization (JMIM) [22] and maximizing independent classification information (MRI) [23]. Since these algorithms consider TDMI among features and the class label

✉ Jichang Guo
jcguo@tju.edu.cn

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

and ignore TDMI among features, their objective functions might miss some useful information and the performance of these algorithms is influenced.

Considering the above problem, this paper investigates feature selection based on TDMI among features. Firstly, to select the features that provide more useful information, based on the maximal relevance minimal redundancy (MRMR) criterion, joint mutual information (JMI) among the class label and feature set as well as MI between feature set is employed to describe relevance and redundancy separately. Then, JMI among the class label and feature set as well as MI between features is decomposed, and TDMI among features is adopted. Furthermore, both performance and computation are considered, and an objective function is achieved. Finally, a feature selection algorithm based on CMI is proposed.

The main contributions of this paper are as follows. (1) The maximal relevance minimal redundancy criterion is adopted in selecting features. (2) Our algorithm takes special consideration of three-dimensional mutual information among features. (3) The proposed algorithm takes both performance and computation into account. (4) Our algorithm can achieve better feature selection performance at the expense of more time-consuming.

The rest of this paper is organized as follows. Section 2 gives the knowledge of mutual information. Related works are analyzed in Section 3. Section 4 presents the proposed algorithm. Experimental results and analysis are given in Section 5. Section 6 is conclusions and future work.

2 The knowledge of mutual information

Assuming x and y are two discrete random variables. $p(x)$ and $p(y)$ are the probability of x and y separately. Information entropy is exploited to measure information and $H(X)$ is defined by (1):

$$H(X) = -\sum_{x\in X} p(x)\log p(x) \tag{1}$$

Conditional entropy $H(Y|X)$ is the entropy of Y when X is given and it can be expressed as (2):

$$H(Y|X) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y|x) \tag{2}$$

where $p(y|x)$ is the conditional probability and $p(x,y)$ is the joint probability.

MI is employed to quantify the information that two variables share and MI $I(X;Y)$ is defined as (3):

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{3}$$

Greater MI value suggests more information that the two variables share. MI has the relationship with information entropy and conditional entropy as (4).

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

TDMI is a supplement of MI and it includes CMI, JMI and three-way interaction information. CMI is the reduction in the uncertainty of the other variable due to the knowledge of another variable when one variable is given. CMI $I(X;Y|Z)$ is defined by (5):

$$I(X;Y|Z) = \sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z} p(x,y,z)\log\frac{p(x,y|z)}{p(x|z)p(y|z)} \tag{5}$$

JMI is utilized to measure the information that two variables share with the other variable. JMI $I(X,Y;Z)$ has the relationship with $I(Y;Z)$ and $I(X;Z|Y)$ as (6).

$$I(X,Y;Z) = I(Y;Z) + I(X;Z|Y) \tag{6}$$

Three-way interaction information $I(X;Y;Z)$ has the relationship with $I(X;Z|Y)$ and $I(X;Z)$ as (7) [24].

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

3 Related works

MIM is a basic feature selection algorithm based on MI and the objective function is presented in (8).

$$MIM = \arg\max_{f_i\in X}\left[I(c;f_i)\right] \tag{8}$$

MIM calculates MI between the class label c and each candidate feature f_i, and selects some features with greater values from the candidate feature set X.

Some feature selection algorithms consider redundancy between features further and algorithms based on relevance and redundancy are proposed, such as mutual information feature selection (MIFS) [25], MIFS-U [26], mRMR, normalized mutual information feature selection (NMIFS) [16], CMI and MIFS-CR. These algorithms exploit MI between candidate features and the class label to describe relevance, and employ MI between selected features and candidate features to describe redundancy. Objective functions are the key of these algorithms. The objective functions of MIFS and MIFS-U are given below.

$$MIFS = \arg\max_{f_i\in X}\left[I(c;f_i) - \beta\sum_{f_s\in S} I(f_s;f_i)\right] \tag{9}$$

$$MIFS{-}U = \arg\max_{f_i \in X}\left[I\left(c;f_i\right) - \beta \sum_{f_s \in S} \frac{I\left(c;\, f_s\right)}{H(f_s)} I\left(f_s;f_i\right)\right] \tag{10}$$

where β is a parameter, f_s is a selected feature and S is the selected feature set.

Since MIFS and MIFS-U have the problem that β is uncertain, mRMR adopts the reciprocal of the number of selected features $|S|$ to replace β and the objective function is shown in (11).

$$mRMR = \arg\max_{f_i \in X}\left[I\left(c;f_i\right) - \frac{1}{|S|} \sum_{f_s \in S} I\left(f_s;f_i\right)\right] \tag{11}$$

Since mRMR is considered to have bias toward the features with greater MI values, MI between a candidate feature and a selected feature is normalized and NMIFS is proposed. Its objective function is given in (12).

$$NMIFS = \arg\max_{f_i \in X}\left[I\left(c;\, f_i\right) - \frac{1}{|S|} \sum_{f_s \in S} \frac{I\left(f_i;\, f_s\right)}{\min\left(H\left(f_i\right), H\left(f_s\right)\right)}\right] \tag{12}$$

CMI and MIFS-CR are proposed, and their objective functions are shown below.

$$CMI = \arg\max_{f_i \in X}\left[I\left(c;f_i\right) - \frac{H(f_i|c)}{H(f_i)} \sum_{f_s \in S} \frac{I(c;\, f_s) I\left(f_s;f_i\right)}{H(f_s)H(c)}\right] \tag{13}$$

$$MIFS{-}CR = \arg\max_{f_i \in X}\left[I\left(c;\, f_i\right) - \frac{1}{2} \sum_{f_s \in S}\left(\frac{I(c;\, f_s)}{H(f_s)} + \frac{I(c;\, f_i)}{H(f_i)}\right) I\left(f_s;\, f_i\right)\right] \tag{14}$$

Furthermore, considering that mRMR has the problem that the feature with the maximum difference is not always the feature with minimal redundancy maximal relevance, a method of equal interval division is adopted to process the case where the objective function of mRMR is not accurate. Finally an algorithm based on equal interval division and minimal-redundancy-maximal-relevance (EID-mRMR) [27] is proposed. Since relevance and redundancy are both considered, these algorithms based on relevance and redundancy can achieve better feature selection performance.

There are some feature selection algorithms based on TDMI, such as dynamic weighting-based feature selection (DWFS) [28], IWFS, conditional mutual information maximization (CMIM) [29], JMI [30], maximal conditional mutual information (MCMI) [31] and MRI. DWFS and IWFS belong to the same category, and they utilize symmetrical uncertainty that normalizes MI to describe relevance.

JMI, CMIM, MCMI, and MRI fall into a different category, and they have different objective functions. Except for the four algorithms above, the kind of algorithms include conditional informative feature extraction (CIFE) [32], JMIM, CFR [33], and Dynamic Change of Selected Feature with the class (DCSF) [34], and their objective functions are presented below.

$$JMI = \arg\max_{f_i \in X}\left[I\left(c;f_i\right) - \frac{1}{|S|} \sum_{f_s \in S} I\left(f_s;f_i\right) + \frac{1}{|S|} \sum_{f_s \in S} I\left(f_s;f_i|c\right)\right] \tag{15}$$

$$CMIM = \arg\max_{f_i \in X}\left[\min_{f_s \in S}\left(I\left(c;\, f_i|f_s\right)\right)\right] \tag{16}$$

$$MCMI = \arg\max_{f_i \in X}\left[\max_{f_s \in S}\left(I\left(c;\, f_i|f_s\right)\right)\right] \tag{17}$$

$$MRI = \arg\max_{f_i \in X}\left[I\left(c;f_i\right) + \sum_{f_s \in S} I\left(c;f_i|f_s\right) + \sum_{f_s \in S} I\left(c;f_s|f_i\right)\right] \tag{18}$$

$$CIFE = \arg\max_{f_i \in X}\left[I\left(c;f_i\right) - \sum_{f_s \in S} I\left(f_s;f_i\right) + \sum_{f_s \in S} I\left(f_s;f_i|c\right)\right] \tag{19}$$

$$JMIM = \arg\max_{f_i \in X}\left[\min_{f_s \in S}\left(I\left(f_i,\, f_s;\, c\right)\right)\right] \tag{20}$$

$$CFR = \arg\max_{f_i \in X}\left[\sum_{f_s \in S} I\left(c;f_i|f_s\right) + \sum_{f_s \in S} I\left(c;f_s;f_i\right)\right] \tag{21}$$

In (15)–(21), since TDMI among features is not employed, feature selection effectiveness of these algorithms can be affected.

4 The proposed algorithm

This section decomposes JMI among feature set and the class label as well as MI between feature set and attains an objective function. Then, based on the objective function, the proposed algorithm is presented.

4.1 The proposed objective function

The aim of some existing feature selection algorithms based on MI and TDMI is to select the feature set that have maximum MI with the class label, and these algorithms only consider relevant information satisfying the maximum. However, selecting a candidate feature f_i introduces not only relevant information, but also redundant information. To introduce maximal relevant and minimal redundant information, we formulate this issue as (22).

$$\arg\max_{f_i \in X} [I(S,f_i;c) - I(S;f_i)] \tag{22}$$

where S is the selected feature set and X is the candidate feature set. The total of relevant information that is introduced is $I(S,f_i;c)$ and that of redundant information is $I(S;f_i)$. The greater the difference between relevant and redundant information, the more informative the candidate feature is. Adopting (22) can ensure that maximal relevant and minimal redundant information is introduced, thus guaranteeing feature selection effectiveness of selected features. $I(S,f_i;c)$ satisfies (23).

$$I(S,f_i;c) = I(S;c) + I(f_i;c|S) \tag{23}$$

$I(f_i;c|S)$ satisfies (24).

$$I(f_i;c|S) = I(c;f_i) - I(S;c) + I(S;c|f_i) \tag{24}$$

Equation (25) is derived by adding (23) to (24).

$$I(S,f_i;c) = I(c;f_i) + I(S;c|f_i) \tag{25}$$

By combining (25) with (22), (26) is obtained.

$$\arg\max_{f_i \in X} [I(c;f_i) + I(S;c|f_i) - I(S;f_i)] \tag{26}$$

$I(S;c|f_i)$ satisfies (27).

$$I(S;c|f_i) = \frac{1}{|S|}\sum_{f_s \in S} I(f_s;c|f_i) \tag{27}$$

By combining (25) with (27), (28) is derived.

$$I(S,f_i;c) = I(c;f_i) + \frac{1}{|S|}\sum_{f_s \in S} I(f_s;c|f_i) \tag{28}$$

$I(S;f_i)$ satisfies (29).

$$I(S;f_i) = \frac{1}{|S|}\sum_{f_s \in S} I(f_i;f_s) + \frac{1}{|S|\,|S-1|}\sum_{f_s \in S}\sum_{f_j \in S, f_j \neq f_s} I(f_j;f_i|f_s) \tag{29}$$

Since it is quite time-consuming to calculate the second part of (29), we replace (29) by (30).

$$\frac{1}{|S|}\sum_{f_s \in S} I(f_i;f_s) + \frac{1}{|S|\,|S-1|}\sum_{f_s \in S}\sum_{f_j \in S, f_j \neq f_s} I(f_j;f_s|f_i) \tag{30}$$

By combining (28) and (30), (31) is obtained.

$$\arg\max_{f_i \in X} \left[I(c;f_i) - \frac{1}{|S|}\sum_{f_s \in S} I(f_i;f_s) + \frac{1}{|S|}\sum_{f_s \in S} I(f_s;c|f_i) - \frac{1}{|S|\,|S-1|}\sum_{f_s \in S}\sum_{f_j \in S, f_j \neq f_s} I(f_j;f_s|f_i) \right] \tag{31}$$

Figure 1 gives a brief description of the determination of (31). Equation (31) considers not only MI between features and the class label, CMI among features and the class label as well as MI between features, but also CMI among features, while the objective functions of other feature selection algorithms do not consider TDMI among features. Compared with the objective functions of other algorithms, (31) contains more useful information and selecting the feature satisfying (31) can guarantee that maximal relevant and minimal redundant information is obtained. To guarantee introducing maximal relevant and minimal redundant information, (31) is taken as an objective function.

4.2 Algorithmic implementation

Based on (31), a feature selection algorithm based on CMI for MRMR (CMI-MRMR) is proposed and the flow chart is presented in Fig. 2.

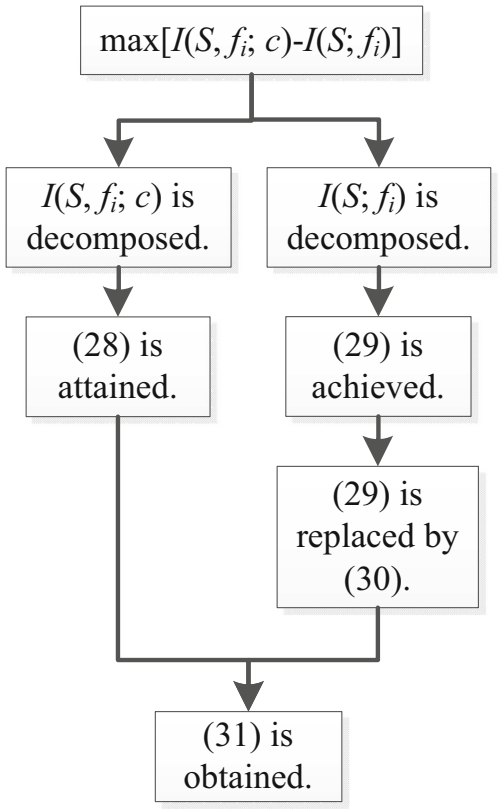

Fig. 1 Determination of (31)

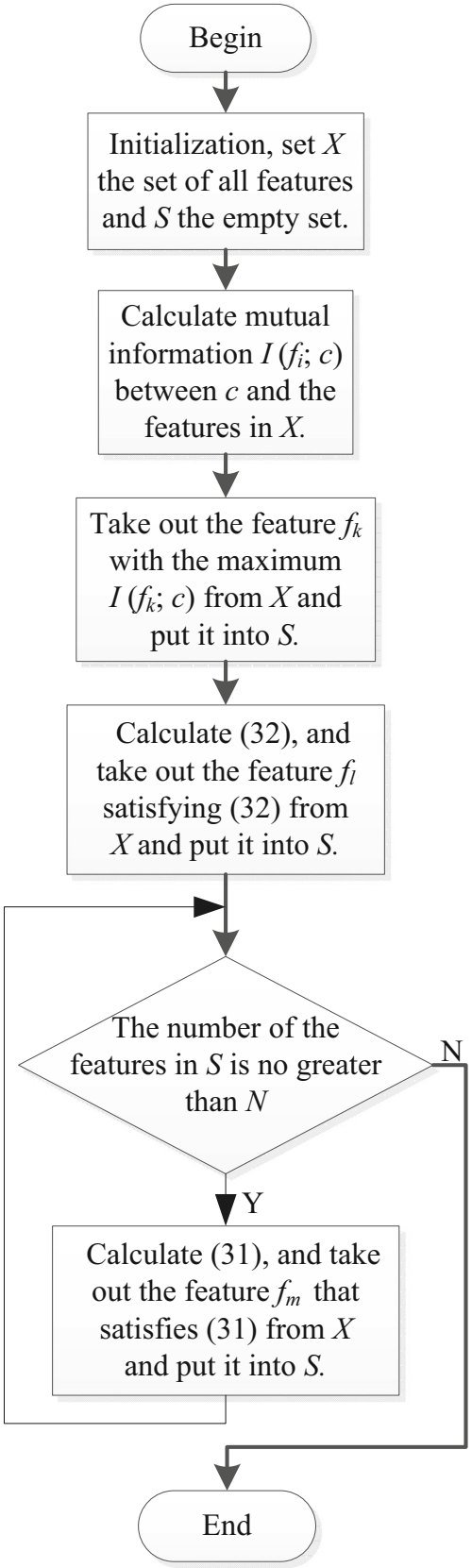

Fig. 2 Flow chart of CMI-MRMR

In Fig. 2, we first initialize X and S. Then, we calculate MI between features and the class label, and select the feature with maximum. We calculate (32) that does not have the fourth part of (31) and select the feature satisfying the condition. Following that, we calculate (31) and select the feature that meets the requirement until a specified number of features are selected.

$$\arg\max_{f_i \in X} \left[I(c;f_i) - \frac{1}{|S|}\sum_{f_s \in S} I\left(f_i;f_s\right) + \frac{1}{|S|}\sum_{f_s \in S} I\left(f_s;c|f_i\right) \right] \tag{32}$$

The pseudo-code of CMI-MRMR is shown in Algorithm 1.

Algorithm 1 A Feature Selection Algorithm based on CMI for MRMR(CMI-MRMR).

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

1: initialize $S = \emptyset$ and $X = \{f_1, f_2, \ldots, f_M\}$.
2: **for** $f_i \in X$ **do**
3: compute $I\left(f_i;c\right)$.
4: **end for**
5: take the feature f_k by maximizing MI from X and put it into S.
6: $f_s = f_k$.
7: **for** $f_i \in X$ **do**
8: compute $I\left(f_i;f_s\right)$.
9: compute $I\left(f_s;c|f_i\right)$.
10: compute (32).
11: **end for**
12: take the feature f_l by satisfying (32) from X and put it into S.
13: **while** $|S| \leq N$ **do**
14: **for** $f_i \in X$ **do**
15: **for** $f_s \in S$ **do**
16: **for** $f_j \in S$ and $f_j \neq f_s$ **do**
17: compute $I\left(f_j;f_s|f_i\right)$.
18: **end for**
19: compute $I\left(f_i;f_s\right)$.
20: compute $I\left(f_s;c|f_i\right)$.
21: **end for**
22: compute (31).
23: **end for**
24: take the feature f_m by satisfying (31) from X and put it into S.
25: **end while**

CMI-MRMR consists of three parts. The first part (lines 1-6), S and X are initialized. Then, MI between the class

Table 1 Description of the datasets

Datasets	Instances	Features	Classes	Types	Sources
Musk	476	166	2	Continuous	UCI
Mfeat_fac	2000	216	10	Continuous	UCI
Mfeat_pix	2000	240	10	Discrete	UCI
Semeion	1593	256	10	Discrete	UCI
USPS	9298	256	10	Continuous	ASU
lung_discrete	73	325	7	Discrete	ASU
Isolet	1560	617	26	Continuous	ASU
COIL20	1440	1024	20	Continuous	ASU
warpAR10P	130	2400	10	Continuous	ASU
lung	203	3312	5	Continuous	ASU
gisette	7000	5000	2	Continuous	ASU
Carcinom	174	9182	11	Continuous	ASU
pixraw10P	100	10000	10	Continuous	ASU
arcene	200	10000	2	Continuous	ASU
orlraws10P	100	10304	10	Continuous	ASU
CLL_SUB_111	111	11340	3	Continuous	ASU

Table 2 Classification accuracy (%) of selected features with J48

Datasets	CMI-MRMR	mRMR	IWFS	JMIM	MCMI	MRI	CFR	DCSF
Musk	80.56±0.73	80.43±0.44	80.31±1.18	78.58±0.86	78.08±0.56	78.73±0.77	78.81±0.71	80.24±0.74
Mfeat_fac	85.74±0.19	85.12±0.24	84.33±0.34	85.15±0.36	82.42±0.19	85.50±0.36	85.11±0.39	85.22±0.25
Mfeat_pix	74.42±0.34	73.25±0.31	71.55±0.37	73.62±0.41	64.48±0.46	73.58±0.38	73.70±0.29	74.25±0.38
Semeion	65.24±0.42	62.88±0.35	60.98±0.59	65.36±0.27	58.64±0.25	63.82±0.26	64.16±0.21	66.82±0.46
USPS	83.46±0.09	81.34±0.12	81.69±0.13	81.16±0.09	72.80±0.05	80.98±0.10	81.00±0.09	84.28±0.13
lung_discrete	45.81±2.89	46.13±3.15	43.91±3.76	45.65±3.38	43.79±1.22	40.57±3.56	40.90±3.60	40.36±1.41
Isolet	68.38±0.57	66.30±0.52	63.71±0.62	64.26±0.40	53.54±0.28	67.15±0.49	67.14±0.49	68.86±0.35
COIL20	89.41±0.29	88.50±0.50	86.53±0.36	88.12±0.35	83.79±0.34	87.72±0.39	86.98±0.47	86.70±0.37
warpAR10P	68.33±1.96	67.83±1.92	58.67±1.47	66.81±1.76	65.91±2.39	67.30±1.60	66.89±1.81	64.94±1.64
lung	89.75±0.71	88.81±0.89	86.16±1.69	88.65±1.30	87.62±0.87	89.43±1.02	88.69±1.06	86.86±1.29
gisette	92.57±0.07	92.01±0.10	92.06±0.21	91.23±0.12	92.38±0.07	92.86±0.07	92.84±0.08	93.37±0.08
Carcinom	71.20±0.81	71.45±1.35	63.21±1.66	69.14±2.67	69.87±1.35	68.61±1.77	67.67±1.88	64.81±2.12
pixraw10P	93.01±1.20	92.83±1.80	89.46±2.83	91.33±2.43	91.59±1.54	89.37±1.46	89.25±2.71	88.62±2.18
arcene	77.45±2.37	76.96±2.17	78.00±1.85	72.60±1.51	73.82±1.96	72.43±1.61	72.51±1.37	74.25±1.07
orlraws10P	75.99±2.18	74.77±2.60	72.30±4.85	74.18±4.60	70.18±2.88	75.62±4.02	75.92±3.51	75.11±2.43
CLL_SUB_111	66.84±3.31	67.74±3.17	65.51±3.01	66.84±2.55	67.20±2.95	65.24±3.27	64.16±2.93	64.37±2.46
Avg.	76.76	76.02	73.65	75.17	72.26	74.93	74.73	74.94
W/T/L	–	8/8/0	12/4/0	12/4/0	15/1/0	10/5/1	14/1/1	9/3/4

Table 3 Classification accuracy (%) of selected features with IB1

Datasets	CMI-MRMR	mRMR	IWFS	JMIM	MCMI	MRI	CFR	DCSF
Musk	82.02±0.52	81.90±0.51	82.23±0.57	79.85±0.49	79.32±0.55	80.03±0.79	80.07±0.84	81.74±0.53
Mfeat_fac	92.16±0.10	91.55±0.14	91.43±0.17	91.43±0.11	88.34±0.19	91.71±0.13	91.47±0.14	91.79±0.14
Mfeat_pix	77.75±0.24	73.77±0.18	69.93±0.40	76.00±0.45	59.93±0.55	75.89±0.14	76.39±0.21	79.23±0.31
Semeion	62.46±0.35	58.89±0.25	57.40±0.51	62.35±0.41	52.49±0.21	60.15±0.26	60.72±0.36	65.50±0.39
USPS	84.81±0.08	82.41±0.06	82.22±0.30	82.15±0.11	67.90±0.07	82.16±0.14	82.26±0.11	87.90±0.11
lung_discrete	71.57±2.48	70.67±2.04	64.11±2.38	70.15±2.08	65.27±2.40	71.75±2.09	70.85±1.76	71.75±2.52
Isolet	71.90±0.67	70.06±0.38	66.53±0.32	67.55±0.46	55.27±0.48	70.80±0.56	70.94±0.60	74.15±0.49
COIL20	95.78±0.12	95.45±0.12	95.46±0.15	95.49±0.13	91.66±0.28	95.93±0.07	95.83±0.16	95.73±0.13
warpAR10P	77.45±1.36	77.15±1.08	69.81±1.74	79.68±1.75	77.91±1.29	77.58±1.14	76.60±1.37	73.46±1.21
lung	92.80±0.31	92.81±0.28	89.37±0.90	92.03±0.63	90.41±0.82	92.66±0.52	92.78±0.38	91.78±0.89
gisette	91.74±0.10	91.30±0.14	91.19±0.17	90.28±0.12	91.50±0.05	92.71±0.06	92.57±0.05	93.05±0.09
Carcinom	82.13±1.10	81.36±1.08	76.30±1.25	81.02±1.34	80.50±1.29	81.45±1.06	81.27±1.19	81.20±1.20
pixraw10P	96.26±0.45	96.34±0.70	86.07±2.02	94.51±1.01	91.34±1.34	94.26±0.95	93.98±0.95	92.68±0.98
arcene	81.47±1.57	81.09±1.05	80.99±1.32	79.91±1.20	78.60±1.21	79.89±1.39	79.71±1.41	81.04±1.12
orlraws10P	90.64±1.20	91.11±1.48	82.36±2.17	90.62±1.87	85.86±1.67	89.89±1.59	89.40±1.17	88.56±1.57
CLL_SUB_111	72.38±1.93	71.94±1.59	66.00±2.12	68.77±2.63	68.25±1.81	71.89±2.13	70.89±1.98	69.60±2.38
Avg.	82.71	81.74	78.21	81.36	76.53	81.80	81.61	82.45
W/T/L	–	8/8/0	14/2/0	13/2/1	15/1/0	9/5/2	12/3/1	8/3/5

Table 4 Classification accuracy (%) of selected features with Naive Bayes

Datasets	CMI-MRMR	mRMR	IWFS	JMIM	MCMI	MRI	CFR	DCSF
Musk	79.06±0.58	79.29±0.41	77.29±0.82	77.87±0.60	75.87±1.03	78.48±0.52	78.46±0.56	78.88±0.38
Mfeat_fac	89.61±0.12	88.43±0.20	87.08±0.24	87.86±0.19	82.70±0.16	88.49±0.20	88.22±0.19	89.41±0.14
Mfeat_pix	81.40±0.17	79.14±0.18	72.68±0.40	79.36±0.24	65.88±0.39	79.05±0.21	79.01±0.13	81.32±0.14
Semeion	65.95±0.15	63.52±0.13	54.67±0.38	64.82±0.14	57.50±0.25	64.35±0.18	64.64±0.16	68.35±0.21
USPS	81.72±0.07	79.81±0.06	74.38±0.07	75.23±0.10	60.38±0.09	76.00±0.15	75.93±0.12	82.41±0.07
lung_discrete	75.81±1.51	73.56±1.90	71.46±2.10	74.45±2.00	68.76±2.16	75.67±1.56	74.79±2.29	76.55±2.23
Isolet	70.67±0.33	67.21±0.43	68.68±0.35	64.69±0.27	50.82±0.46	70.04±0.44	70.25±0.39	73.68±0.33
COIL20	93.74±0.18	92.00±0.20	90.71±0.18	91.70±0.32	81.84±0.81	93.27±0.25	92.69±0.42	92.69±0.39
warpAR10P	70.69±1.13	69.97±1.29	67.14±1.95	66.64±1.71	64.08±1.59	69.04±1.72	68.74±1.58	71.29±1.75
lung	93.26±0.29	92.72±0.56	91.05±0.66	92.32±0.61	88.70±1.20	92.88±0.74	92.76±0.79	92.34±1.15
gisette	88.65±0.06	88.28±0.03	86.18±0.23	86.03±0.05	84.98±0.04	87.60±0.03	87.46±0.04	89.51±0.06
Carcinom	80.44±0.76	79.26±0.79	76.31±1.70	79.22±1.02	78.13±1.31	79.90±1.10	79.66±1.08	80.16±1.21
pixraw10P	97.33±0.65	97.05±0.57	86.50±1.55	95.90±0.46	89.57±1.62	97.22±0.53	96.91±0.73	95.23±0.82
arcene	73.51±1.30	72.66±1.30	72.50±1.27	69.76±0.54	69.25±1.40	70.24±0.68	70.38±0.65	74.52±1.05
orlraws10P	92.27±1.28	90.85±1.01	78.12±2.83	89.97±1.21	77.88±1.33	92.06±1.62	90.89±1.77	90.53±1.66
CLL_SUB_111	76.48±2.15	76.31±2.15	70.95±2.66	73.37±2.23	70.95±3.14	75.12±1.80	75.27±2.46	73.05±2.64
Avg.	81.91	80.63	76.61	79.32	72.96	80.59	80.38	81.87
W/T/L	–	13/3/0	16/0/0	16/0/0	16/0/0	13/3/0	14/2/0	7/4/5

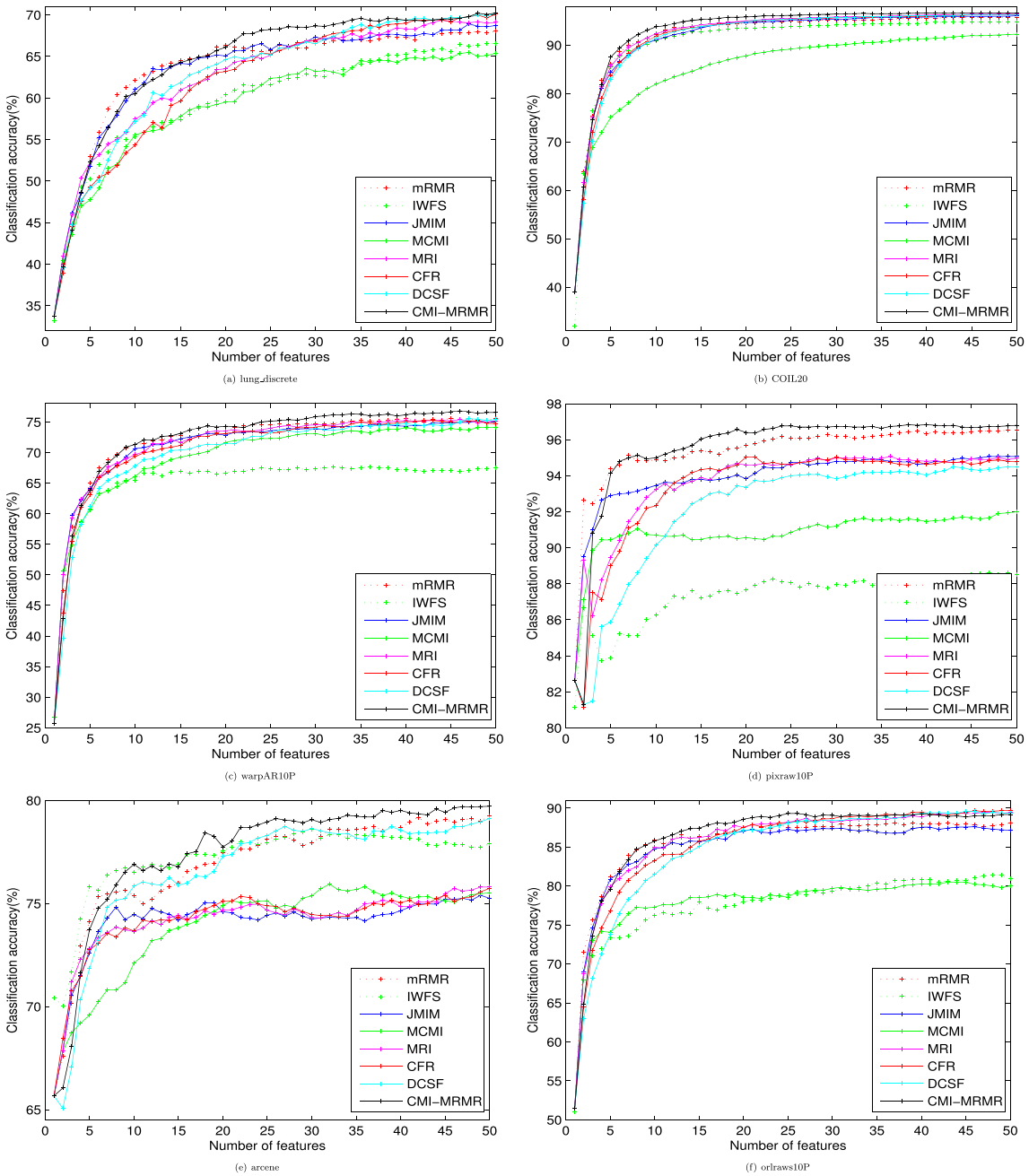

Fig. 3 Average performance comparisons of algorithms with the three classifiers

label and features is calculated, and the feature f_k is selected; in the second part (lines 7-12), $I(f_i;f_s)$ and $I(f_s;c|f_i)$ are calculated. Then, (32) is calculated and the feature f_l that satisfies the condition is selected from X; in the third part (lines 13-25), $I\left(f_j;f_s|f_i\right)$, $I(f_s;c|f_i)$, and $I(f_i;f_s)$ are calculated. Then, (31) is calculated and the feature f_m meeting the requirement is selected. Following the above steps, the process ends when the number of selected features is N.

5 Experimental results

To validate the performance of CMI-MRMR, mRMR, IWFS, JMIM, MCMI, MRI, CFR, and DCSF are compared.

5.1 The datasets and experimental settings

The datasets in Table 1 are from UCI machine learning repository [35] and Arizona State University (ASU) feature

Table 5 Classification accuracy (%) of the optimal features selected by CMI-MRMR and all features

Datasets	J48		IB1		Naive Bayes	
	CMI-MRMR	All features	CMI-MRMR	All features	CMI-MRMR	All features
Musk	**83.37**	82.94	84.90	**85.35**	**82.32**	81.84
Mfeat_fac	**88.83**	88.74	**96.41**	95.96	**93.78**	93.13
Mfeat_pix	78.38	**78.50**	91.28	**95.93**	88.48	**93.46**
Semeion	71.89	**75.81**	78.71	**91.48**	75.76	**85.46**
USPS	88.02	**89.37**	93.38	**97.32**	**85.29**	85.11
lung_discrete	**48.07**	43.86	81.09	**82.70**	84.14	**88.80**
Isolet	75.06	**79.19**	81.44	**90.15**	78.47	**89.62**
COIL20	92.40	**93.43**	99.67	**99.90**	98.28	**98.35**
warpAR10P	**71.85**	70.54	**82.54**	49.23	76.54	**78.23**
lung	90.57	**92.10**	**94.33**	94.20	94.63	**95.96**
gisette	**94.10**	93.94	94.22	**96.16**	89.61	**90.85**
Carcinom	**75.63**	75.49	**89.22**	85.36	88.79	**90.11**
pixraw10P	**94.30**	92.90	98.30	**98.80**	**98.70**	97.80
arcene	**79.60**	74.45	84.95	**85.95**	**75.70**	67.25
orlraws10P	**78.30**	69.00	**95.00**	93.70	96.80	**97.70**
CLL_SUB_111	**68.92**	61.11	**75.48**	63.70	**81.09**	74.07
Avg.	**79.96**	78.84	**88.81**	87.87	86.77	**87.98**

Table 6 The number of the optimal features selected by CMI-MRMR and all features

Datasets	CMI-MRMR			All features
	J48	IB1	Naive Bayes	
Musk	40	44	49	166
Mfeat_fac	50	41	33	216
Mfeat_pix	47	50	49	240
Semeion	50	50	50	256
USPS	50	50	34	256
lung_discrete	14	48	50	325
Isolet	50	50	50	617
COIL20	33	50	50	1024
warpAR10P	46	49	46	2400
lung	47	42	50	3312
gisette	41	34	39	5000
Carcinom	34	50	50	9182
pixraw10P	24	50	19	10000
arcene	22	40	49	10000
orlraws10P	30	50	49	10304
CLL_SUB_111	25	50	46	11340
Avg.	37.69	46.75	44.56	4039.88

selection datasets [36]. For all the datasets, N is set to 50. Minimum description length discretization method [37] is exploited to transform the numerical features into discrete ones. Three popular classifiers, J48, IB1, and Naive Bayes are employed and their parameters are set to Waikato environment for knowledge analysis (WEKA)'s [38] default values. ASU feature selection software package [39] is utilized.

5.2 Experimental results and analysis

To reduce the influence of randomness on the final results, ten times of 10-fold cross-validation are employed, and the mean value and standard deviation of ten results are taken as the final results. Classification accuracy of features selected by these algorithms with J48, IB1 and Naive Bayes is presented in Tables 2–4. To determine whether the effectiveness of experimental results is significant, a one-sided paired t-test at 5% significance level is performed, and the number of the datasets that CMI-MRMR performs better than/equal to/worse than other algorithms is shown in Win/Tie/Loss (W/T/L). Average performance of algorithms with the three classifiers is given in Fig. 3. Furthermore, the optimal top several features from 1 to 50 selected by CMI-MRMR is compared with all features, and the comparison result with three classifiers is shown in Tables 5 and 6.

As shown in Table 2, the Avg. values show that mRMR, JMIM and CMI-MRMR achieve better results. For the W/T/L values, the number of datasets that mRMR, DCSF and CMI-MRMR can obtain better feature selection performace.

In Table 3, mRMR, DCSF and CMI-MRMR obtain greater Avg. and W/T/L values with IB1. In comparison with Table 2, CMI-MRMR outperforms mRMR, IWFS and MCMI with more performance gain in the Avg. values and it achieves more advantages than IWFS in the W/T/L values.

The Avg. and W/T/L values in Table 4 show that mRMR, MRI, DCSF and CMI-MRMR obtain better feature selection effectiveness. Compared with Tables 2 and 3, in terms of the Avg. values, CMI-MRMR has more advantage than mRMR, IWFS, JMIM, and MCMI. For the W/T/L values, CMI-MRMR can obtain better performance gain than other algorithms except DCSF.

As shown in Fig. 3, CMI-MRMR achieves better feature selection effectiveness in the majority of datasets, while other algorithms cannot obtain the desired results in some datasets. We take mRMR and CFR as examples, mRMR cannot handle well in lung_discrete and orlraws10P. CFR does not achieve the desired feature selection performance in lung_discrete and arcene.

As shown in Tables 5 and 6, the Avg. values show that the optimal features selected by CMI-MRMR obtain higher accuracy than all features. In comparison with these datasets, the number of the datasets that the features selected by CMI-MRMR perform better than all features is 8 with J48, 4 with IB1, and 5 with Naive Bayes. Overall, although CMI-MRMR only selects the top 50 features, it can have fairly good performance with all features.

6 Conclusions and future work

This paper investigates feature selection based on TDMI among features and proposes a feature selection algorithm named CMI-MRMR. To verify the performance, we apply it to three classifiers, four UCI datasets, and twelve ASU datasets, and compare results with those from several algorithms based on MI and TDMI. Experimental results validate that CMI-MRMR can achieve better feature selection effectiveness. Furthermore, the optimal feature set selected by CMI-MRMR are compared with all features, the comparison results show that CMI-MRMR can achieve fairly good performance with all features in the majority of datasets, even better than all features in some datasets.

Considering that CMI-MRMR can achieve better feature selection performance, it can be applied in many fields, such as text processing, underwater objects recognition and classification, network anomaly detection, gene expression, and image classification. Classification accuracy of the top optimal several features from 1 to 50 selected by CMI-MRMR is compared with all features, since classification results of the top optimal several features are worse than all features in some datasets, the determination of the number of selected features will be investigated in the next stage.

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. Brown G, Pocock A, Zhao MJ, Lujan M (2012) Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. J Mach Learn Res 13:27–66
3. Bolon CV, Sanchez MN, Alonso BA (2015) Recent advances and emerging challenges of feature selection in the context of big data. Knowl.-Based Syst 86:33–45
4. 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
5. 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
6. 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

7. 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
8. Gu XY, Guo JC (2019) A study on Subtractive Pixel Adjacency Matrix features. Multimed Tools Appl 78(14):19681–19695
9. 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
10. 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
11. Zhang F, Chan PPK, Biggio B, Yeung DS, Roli F (2016) Adversarial feature selection against evasion attacks. IEEE Trans Cybern 46(3):766–777
12. Veronica BC, Noelia SM, Amparo AB (2013) A review of feature selection methods on synthetic data. Knowl Inf Syst 34(3):483–519
13. Jia XP, Kuo BC, Crawford MM (2013) Feature Mining for Hyperspectral Image Classification. Proc IEEE 101(3):676–697
14. Lin CH, Chen HY, Wu YS (2014) Study of image retrieval and classification based on adaptive features using genetic algorithm feature selection. Expert Syst Appl 41(15):6611–6621
15. Naghibi T, Hoffmann S, Pfister B (2015) A Semidefinite Programming Based Search Strategy for Feature Selection with Mutual Information Measure. IEEE Trans Pattern Anal Mach Intell 37(8):1529–1541
16. Estevez PA, Tesmer M, Perez CA, Zurada JA (2009) Normalized Mutual Information Feature Selection. IEEE Trans Neural Netw 20(2):189–201
17. Lewis DD (1992) Feature selection and feature extraction for text categorization. In: Proceedings of the workshop on speech and natural language, pp 212–217
18. 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
19. 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
20. Wang ZC, Li MQ, Li JZ (2015) A multi-objective evolutionary algorithm for feature selection based on mutual information with a new redundancy measure. Inform Sci 307:73–88
21. Zeng ZL, Zhang HJ, Zhang R, Yin CX (2015) A novel feature selection method considering feature interaction. Pattern Recogn 48(8):2656–2666
22. Bennasar M, Hicks Y, Setchi R (2015) Feature selection using Joint Mutual Information Maximisation. Expert Syst Appl 42(22):8520–8532
23. 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
24. Jakulin A, Bratko I (2004) Testing the significance of attribute interactions. In: Proceedings of international conference on machine learning, pp 409–416
25. Battiti R (1994) Using mutual information for selecting features in supervised neural net learning. IEEE Trans Neural Netw 5(4):537–550
26. Kwak N, Choi CH (2002) Input feature selection for classification problems. IEEE Trans Neural Netw 13(1):143–159
27. 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
28. 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
29. Fleuret F (2004) Fast binary feature selection with conditional mutual information. J Mach Learn Res 5:1531–1555
30. Yang HH, Moody JE (1999) Data visualization and feature selection: new algorithms for nongaussian data. In: Proceedings of conference on neural information processing systems
31. Ren JF, Jiang XD, Yuan JS (2015) Learning LBP structure by maximizing the conditional mutual information. Pattern Recogn 48(10):3180–3190
32. Lin DH, Tang X (2006) Conditional infomax learning: An integrated framework for feature extraction and fusion. In: Proceedings of european conference on computer vision. pp 68–82
33. Gao WF, Hu L, Zhang P, He JL (2018) Feature selection considering the composition of feature relevancy. Pattern Recogn Lett 112:70–74
34. Gao WF, Hu L, Zhang P (2018) Class-specific mutual information variation for feature selection. Pattern Recogn 79:328–339
35. Dua D, Graff C (2019) UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences
36. 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)
37. 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
38. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. ACM SIGKDD Explor Newslett 11(1):10–18
39. Zhao Z, Morstatter F, Sharma S, Alelyani S, Anand A, Liu H (2010) Advancing feature selection research. ASU feature selection repository 1–28

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.

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.

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.