Feature Selection by Maximizing Independent Classification Information

Jun Wang, Jin-Mao Wei, Zhenglu Yang, and Shu-Qin Wang

Abstract—Feature selection approaches based on mutual information can be roughly categorized into two groups. The first group minimizes the redundancy of features between each other. The second group maximizes the new classification information of features providing for the selected subset. A critical issue is that large new information does not signify little redundancy, and vice versa. Features with large new information but with high redundancy may be selected by the second group, and features with low redundancy but with little relevance with classes may be highly scored by the first group. Existing approaches fail to balance the importance of both terms. As such, a new information term denoted as *Independent Classification Information* is proposed in this paper. It assembles the newly provided information and the preserved information negatively correlated with the redundant information. Redundancy and new information are properly unified and equally treated in the new term. This strategy helps find the predictive features providing large new information and little redundancy. Moreover, independent classification information is proved as a loose upper bound of the total classification information of feature subset. Its maximization is conducive to achieve a high global discriminative performance. Comprehensive experiments demonstrate the effectiveness of the new approach.

Index Terms—Feature selection, mutual information, feature redundancy, independent classification information

1 Introduction

Mutual information is an effective criterion to measure variable correlation [1]. The mutual information between two variables $\mathbf{y}$ and $\mathbf{x}$ is defined as

$$I(\mathbf{y};\mathbf{x}) = H(\mathbf{y}) - H(\mathbf{y}|\mathbf{x}), \tag{1}$$

where $H(\mathbf{y})$ and $H(\mathbf{y}|\mathbf{x})$ represent the entropy and conditional entropy of the involved variables. It describes the decreased uncertainty for one variable when another variable is given, that is, their shared information [2].

Mutual information is widely utilized to evaluate the discriminative performance of features [3], [4]. These methods aim to find the most relevant features [5], [6] to the target class [7]. This mechanism can be denoted as the maximization of Eq. (2), supposing features $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ are evaluated and $\mathbf{y}$ is the target class for recognition:

$$I(\mathbf{y};\mathbf{x}_1,\ldots,\mathbf{x}_k) = H(\mathbf{y}) - H(\mathbf{y}|\mathbf{x}_1,\ldots,\mathbf{x}_k). \tag{2}$$

- *J. Wang and J.-M. Wei are with the Institute of Big Data, College of Computer and Control Engineering, Nankai University, Tianjin 300071, China. E-mail: junwang@mail.nankai.edu.cn, weijm@nankai.edu.cn.*
- *Z. Yang is with the Institute of Big Data, College of Computer and Control Engineering, and the Institute of Statistics, Nankai University, Tianjin 300071, China. E-mail: yangzl@nankai.edu.cn.*
- *S.-Q. Wang is with the College of Computer and Information Engineering, Tianjin Normal University, Tianjin 300387, China. E-mail: wangsq562@nenu.edu.cn.*

Manuscript received 10 July 2016; revised 16 Dec. 2016; accepted 3 Jan. 2017. Date of publication 10 Jan. 2017; date of current version 3 Mar. 2017. Recommended for acceptance by D. Cai.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2017.2650906

Theoretically, it can be calculated as

$$I(\mathbf{y};\mathbf{x}_1,\ldots,\mathbf{x}_k) = \sum_{\mathbf{y}}\sum_{\mathbf{x}_1}\cdots\sum_{\mathbf{x}_k} p(\mathbf{y},\mathbf{x}_1,\ldots,\mathbf{x}_k) log \frac{p(\mathbf{y},\mathbf{x}_1,\ldots,\mathbf{x}_k)}{p(\mathbf{y})p(\mathbf{x}_1,\ldots,\mathbf{x}_k)}. \tag{3}$$

The features maximizing Eq. (2) are recognized as most discriminative for $\mathbf{y}$ because of their maximal information for classification. An inevitable problem is that joint probabilities in Eq. (2) are complicated to be estimated accurately, unless all of the involved variables are independent identically distributed (i.i.d.) [8]. This issue becomes more intractable on small samples in high dimensions. Even if these joint probabilities can be obtained, an exhaustive search of selecting k optimal features from d candidates is near $O(d^k)$, which is almost impractical for high-dimensional learning tasks [9]. Therefore, mutual information-based methods commonly adopt successive strategies to search for features [10].

These methods can be generally categorized into two groups according to their different feature evaluation criteria, i.e., feature redundancy minimization and newly provided classification information maximization. The methods in the first group focus on reducing feature redundancy to achieve high global classification performance, which is implemented by maximizing the criterion J_{red} as follows:

$$J_{red}(\mathbf{x}_k) = I(\mathbf{y};\mathbf{x}_k) - R(\mathbf{S};\mathbf{x}_k), \tag{4}$$

where $\mathbf{S}$ is the selected feature subset, and $R(\mathbf{S};\mathbf{x}_k)$ represents the redundant information of $\mathbf{x}_k$ with $\mathbf{S}$. Typically, $R(\mathbf{S};\mathbf{x}_k)$ is also computed through mutual information. In some methods [11], [12], [13], it is specifically computed as

$$R(\mathbf{S};\mathbf{x}_k) = \alpha \sum_{\mathbf{x}_j \in \mathbf{S}} I(\mathbf{x}_j;\mathbf{x}_k). \tag{5}$$

Directly using $I(\mathbf{x}_j;\mathbf{x}_k)$ to measure feature redundancy focuses too much on the relevance of $\mathbf{x}_k$ with the selected features to differentiate whether this relevance makes sense to recognizing the target class. Therefore, other methods [14], [15], [16], [17], [18], [19] have been applied to alleviate feature redundancy through multi-information $I(\mathbf{y};\mathbf{x}_j;\mathbf{x}_k)$ [20], which is defined as

$$I(\mathbf{y};\mathbf{x}_j;\mathbf{x}_k) = I(\mathbf{y};\mathbf{x}_j) + I(\mathbf{y};\mathbf{x}_k) - I(\mathbf{y};\mathbf{x}_j,\mathbf{x}_k). \quad (6)$$

Multi-information represents the amount of interaction information corresponding to all referred variables [21]. In the examination of the relationship between two features and classes, this information can be regarded as the shared discriminative information of two features. Given one of the features is selected, multi-information provided by another feature is hence redundant for classification.

In the second group, feature selection methods intend to maximize the new classification information provided for the selected subset by a candidate feature [22], [23], [24], [25]. The criterion J_{new} can be generally defined for these methods as

$$J_{new}(\mathbf{x}_k) = C(\mathbf{y};\mathbf{x}_k|\mathbf{S}). \quad (7)$$

For practical concerns, $J_{new}(\mathbf{x}_k)$ can be computed as

$$J_{new}(\mathbf{x}_k) = \beta \sum_{\mathbf{x}_j \in \mathbf{S}} I(\mathbf{y};\mathbf{x}_k|\mathbf{x}_j). \quad (8)$$

Compared with the criterion of the first group, this criterion is more direct and interpretable. $I(\mathbf{y};\mathbf{x}_k|\mathbf{x}_j)$ quantifies the new classification information provided by $\mathbf{x}_k$ when $\mathbf{x}_j$ is selected. A feature should be highly scored if it provides large new classification information.

Theoretically, for a candidate feature $\mathbf{x}_k$ and a selected feature $\mathbf{x}_j$, we simply have

$$I(\mathbf{y};\mathbf{x}_k|\mathbf{x}_j) = I(\mathbf{y};\mathbf{x}_k) - I(\mathbf{y};\mathbf{x}_j;\mathbf{x}_k). \quad (9)$$

It means that more new information provided by a candidate feature corresponds to less redundancy. This property facilitates transforming some feature selection methods in one group to other methods in another group [26]. Thus, a general framework [26] for mutual information-based methods is formulated as

$$J_{CMI}(\mathbf{x}_k) = I(\mathbf{y};\mathbf{x}_k) - \lambda \sum_{\mathbf{x}_j \in \mathbf{S}} I(\mathbf{x}_k;\mathbf{x}_j) + \gamma \sum_{\mathbf{x}_j \in \mathbf{S}} I(\mathbf{x}_k;\mathbf{x}_j|\mathbf{y}). \quad (10)$$

Within this framework, some mutual information-based methods can be regarded as specialized by adopting different λ and γ. In other words, these methods consider new classification information or redundancy at different degrees. Then, one may wonder which method takes a proper parameter configuration, i.e., appropriately applies complementarity and redundancy in feature evaluation.

Investigating what are considered in existing methods helps find the answer to this problem. Although features are evaluated by different methods from various aspects, these methods concentrate all their attentions on whether a new candidate feature is competitive for classification, but disregard the possible effects on the selected features when a new feature is added. In candidate feature evaluation, its concerned information includes redundant information and newly provided classification information. Correspondingly, the information of a selected feature can also be partitioned into two parts, namely, redundant information and preserved classification information if a new feature is selected. This preserved information varies when different new features are added. Intuitively, candidate features possess little redundancy when selected features preserve large classification information. In other words, the preserved information is negatively correlated with the redundant information. This issue is critical for solving the problem of focusing only on the capabilities of candidate features by existing methods.

To illustrate this problem, we review how a candidate feature is evaluated by existing methods. Methods in the first and second groups respectively focus on redundancy alleviation and new information promotion. According to Eq. (9), it seems that these two kinds of information have been appropriately considered when a candidate feature is examined independently. However, when discussing different candidate features, providing large new classification information does not necessarily mean considering little redundancy, and vice versa. In fact, it is not rare in feature selection that some features strongly relevant to classes are also highly redundant between each other. What often troubles practitioners who use a top-k strategy in feature selection is that the top-ranked features, which are strongly relevant to classes and hence informative and predictive, are likely redundant between each other. Similarly, lowly redundant features with low relevance to classes also have the opportunities to win out under the criterion of alleviating redundancy. This leads to a dilemma about how to achieve balance between feature complementarity and feature redundancy if only examining candidate features. This issue is further elaborated by the following examples.

Real-World Example. A UCI [27] data set named as Haberman's survival data is exploited to illustrate this feature selection problem. It describes the survival status of 306 breast cancer patients at the University of Chicago's Billings Hospital between 1958 and 1970. Three features are involved, i.e., age of patient ($\mathbf{x}_1$), year of operation ($\mathbf{x}_2$), and number of detected positive axillary nodes ($\mathbf{x}_3$). A well-known mutual information Matlab toolbox [28] is used to compute the relevance of features with the target class and redundancy between each other, which is shown as follows:

- $I(\mathbf{y};\mathbf{x}_1) = 0.11, I(\mathbf{y};\mathbf{x}_2) = 0.04, I(\mathbf{y};\mathbf{x}_3) = 0.15$
- $I(\mathbf{x}_1;\mathbf{x}_2) = 1.18, I(\mathbf{x}_1;\mathbf{x}_3) = 1.31, I(\mathbf{x}_2;\mathbf{x}_3) = 0.59$
- $I(\mathbf{y};\mathbf{x}_1|\mathbf{x}_3) = 0.44, I(\mathbf{y};\mathbf{x}_2|\mathbf{x}_3) = 0.32$

Suppose a greedy search strategy is implemented. Thus, the most relevant feature $\mathbf{x}_3$ prevails. Then, selection methods in the two groups draw different conclusions about whether $\mathbf{x}_1$ or $\mathbf{x}_2$ wins out next. According to Eq. (4), the first group selects $\mathbf{x}_2$, that is, the less redundant feature ($J_{red}(\mathbf{x}_2) > J_{red}(\mathbf{x}_1)$). According to Eq. (7), the second group chooses $\mathbf{x}_1$, that is, the more informative feature ($J_{new}(\mathbf{x}_1) > J_{new}(\mathbf{x}_2)$). Clearly, it is still an open question for selection methods to balance the importance of new classification information and redundant information appropriately.

Artificial Example. A group of artificial data set is constructed as follows:

$\mathbf{x}_1$	1	1	1	1	0	0	0	0
$\mathbf{x}_2$	1	1	1	1	0	1	0	0
$\mathbf{x}_3$	0	1	0	1	0	1	0	0
$\mathbf{y}$	0	0	0	0	1	1	1	1

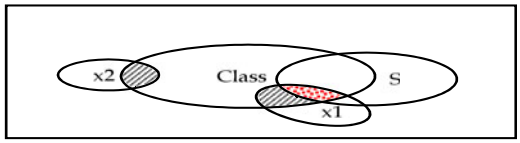


Fig. 1. Relationship between two features and the selected subset S in recognizing the target class: the shadowed parts represent the new classification information, and the red point part represents the class-relevant redundancy.

We observe that $I(\mathbf{y};\mathbf{x}_1) > I(\mathbf{y};\mathbf{x}_2) > I(\mathbf{y};\mathbf{x}_3)$. An interesting situation is that $H(\mathbf{y}|\mathbf{x}_1) = 0$. Thus, $I(\mathbf{y};\mathbf{x}_2|\mathbf{x}_1) = I(\mathbf{y};\mathbf{x}_3|\mathbf{x}_1) = 0$. It indicates that none of the other two features provide any new classification information, once the most relevant feature $\mathbf{x}_1$ is selected. As such, methods in the second group experience difficulty in selecting between $\mathbf{x}_2$ and $\mathbf{x}_3$. As to the methods in the first group, it is also hard for them to make a decision, although they are expected to be able to correctly select the less redundant feature $\mathbf{x}_3$ ($I(\mathbf{x}_1;\mathbf{x}_2) > I(\mathbf{x}_1;\mathbf{x}_3)$). $I(\mathbf{y};\mathbf{x}_2) - I(\mathbf{x}_1;\mathbf{x}_2) = I(\mathbf{y};\mathbf{x}_3) - I(\mathbf{x}_1;\mathbf{x}_3) = 0$ when the parameter α in Eq. (5) is set to 1. Therefore, the first group is also trapped in a dilemma. If $0 \leq \alpha < 1$, the first group incorrectly chooses $\mathbf{x}_2$ rather than $\mathbf{x}_3$. Note that according to Eq. (9), the methods that reduce multi-information redundancy in the first group also experience selection problems in this case.

In fact, this example can be further elaborated by Fig. 1. Suppose $I(\mathbf{y};\mathbf{x}_1|\mathbf{S}) = I(\mathbf{y};\mathbf{x}_2)$, i.e., two shadowed parts in Fig. 1 have equal areas. Then, it is obtained that $I(\mathbf{y};\mathbf{x}_1|\mathbf{S}) = I(\mathbf{y};\mathbf{x}_2|\mathbf{S})$ as well as $I(\mathbf{y};\mathbf{x}_1) - I(\mathbf{y};\mathbf{x}_1;\mathbf{S}) = I(\mathbf{y};\mathbf{x}_2) - I(\mathbf{y};\mathbf{x}_2;\mathbf{S})$. Clearly, both groups face the selection problem, i.e., whether to choose the more predictive feature $\mathbf{x}_1$ or the less redundant feature $\mathbf{x}_2$.

In view of the above analysis, a new information term, *Independent Classification Information* (*ICI*), is proposed in this paper. It unifies redundancy information and new classification information in one term. Thus, the importance of these two kinds of information is synthetically considered by *ICI*. Two kinds of conditional mutual information are employed by *ICI* to evaluate the contributions of candidate and selected features for classification. One kind of information is newly provided by a candidate feature, which denotes the particular contribution of this feature different from that of the selected features. Another kind of information is preserved by the selected features if a candidate feature is selected. This information represents the particular contributions of these features that is different from the candidate feature, and exhibits a negative correlation with the feature redundancy for classification. Therefore, *ICI* focuses on the differences between features in their classification abilities. This strategy helps find highly discriminative as well as lowly redundant features. *ICI* is also proved as a loose upper bound of the global classification information of feature subset. Thus, the new method is expected to obtain a high global classification performance.

The remaining parts of this paper are arranged as follows. Fundamental knowledge and evaluation criteria regarding mutual information are reviewed in Section 2. The concept of independent classification information is introduced in Section 3. A new feature selection criterion is introduced in Section 4 on the basis of this information term. Its several properties are also described in Section 4. Some experimental comparisons of the new method with several popular feature selection methods are presented in Section 5. Conclusions are drawn in Section 6.

2 Related Work

Feature selection is a critical technology to reduce dimensionality. It helps prevent the curse of dimensionality and extract a good representation of the original variable model. Selection methods are typically divided into supervised, semi-supervised, and unsupervised [29]. Supervised methods employ class labels to measure the discriminative abilities of features. This group includes many popular methods, such as *Laplacian Score* [30], *Inf-FS* [31], *ReliefF* [32], just to name a few. There exist some other effective methods that are capable of tackling both supervised and unsupervised tasks, such as *SPEC* [33], *SPFS* [34], and etc. In all related work, including the mutual information-based methods, how to select informative features while reducing feature redundancy is an important issue to be addressed all along.

Intuitively, mutual information can be directly applied to feature selection by maximizing the relevance of candidate feature $\mathbf{x}_k$ with classes, which is represented by the *Max-Relevance* criterion as follows:

$$J_{Max\_Rel}(\mathbf{x}_k) = I(\mathbf{y};\mathbf{x}_k). \tag{11}$$

Discriminative but redundant features are selected by *Max-Relevance*, and thus result in inferior performance to the expected outcome in the recognition task. Therefore, the issue of alleviating redundant information receives more attention [35], [36]. Two representative methods, namely, *MIFS* [11] and *mRMR* [12], are proposed as follows, supposing the feature subset $\mathbf{S} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{k-1}\}$ is selected

$$J_{MIFS}(\mathbf{x}_k) = I(\mathbf{y};\mathbf{x}_k) - \beta \sum_{\mathbf{x}_j \in \mathbf{S}} I(\mathbf{x}_j;\mathbf{x}_k), \tag{12}$$

$$J_{mRMR}(\mathbf{x}_k) = I(\mathbf{y};\mathbf{x}_k) - \frac{1}{|\mathbf{S}|} \sum_{\mathbf{x}_j \in \mathbf{S}} I(\mathbf{x}_j;\mathbf{x}_k). \tag{13}$$

Feature redundancy is reduced by both methods, in which the mutual information of two features is directly considered as their redundancy and minimized.

$I(\mathbf{x}_j;\mathbf{x}_k)$ quantifies the amount of information that two features share, which may or may not be relevant to classification. Obviously, only the information shared by two features to recognize class $\mathbf{y}$ should be regarded as redundant for classification. This information is de facto the multi-information $I(\mathbf{y};\mathbf{x}_j;\mathbf{x}_k)$ in Eq. (6). $I(\mathbf{y};\mathbf{x}_j;\mathbf{x}_k)$ can also be computed as $I(\mathbf{y};\mathbf{x}_j;\mathbf{x}_k) = I(\mathbf{y};\mathbf{x}_k) - I(\mathbf{y};\mathbf{x}_k|\mathbf{x}_j)$ [26]. This implies that information provided by $\mathbf{x}_k$ partially contributes to classification, because this information also involves the redundant information possessed by the selected feature $\mathbf{x}_j$. Note that $I(\mathbf{y};\mathbf{x}_j;\mathbf{x}_k)$ may obtain both positive and negative values [37]. It is positive if adding the condition feature $\mathbf{x}_j$ reduces the relevance of $\mathbf{x}_k$ with $\mathbf{y}$, which can be interpreted as the class-relevant redundancy of two features. Conversely, a negative value is obtained if adding $\mathbf{x}_j$ helps enhance this relevance. In this case, two features are complementary for recognition [38]. Some methods, such as *CIFE* [14], *MIFS-U* [15], *CMIFS* [16], *ICAP* [17], *mIMR* [18], and *IGFS* [19], employ multi-information in their evaluation criteria to determine the redundancy of two features. The criteria of *CIFE* and *ICAP* are shown as follows:

$$J_{CIFE}(\mathbf{x}_k) = I(\mathbf{y};\mathbf{x}_k) - \sum_{\mathbf{x}_j \in \mathbf{S}} I(\mathbf{y};\mathbf{x}_j;\mathbf{x}_k). \quad (14)$$

$$J_{ICAP}(\mathbf{x}_k) = I(\mathbf{y};\mathbf{x}_k) - \sum_{\mathbf{x}_j \in \mathbf{S}} \max[0, I(\mathbf{y};\mathbf{x}_j;\mathbf{x}_k)]. \quad (15)$$

Reducing redundancy can enhance the discriminative ability of a feature subset. A more direct way is to maximize the classification information newly provided for feature subset by candidate features. The joint mutual information between the subset and classes is expected to be increased by this strategy. *JMI* [22], *IF* [23], *DISR* [24], [39], and *CMIM* [25] can be included into this group. In contrast to redundancy reduction methods, which take the target $\mathbf{y}$ as a condition, the selected features are considered as conditions in these methods. *JMI* in Eq. (16) and *CMIM* in Eq. (17) illustrate this idea:

$$J_{JMI}(\mathbf{x}_k) = \sum_{\mathbf{x}_j \in \mathbf{S}} I(\mathbf{x}_k, \mathbf{x}_j; \mathbf{y}) \propto \sum_{\mathbf{x}_j \in \mathbf{S}} I(\mathbf{y};\mathbf{x}_k|\mathbf{x}_j). \quad (16)$$

$$J_{CMIM}(\mathbf{x}_k) = \min_{\mathbf{x}_j \in \mathbf{S}} \big[I(\mathbf{y};\mathbf{x}_k|\mathbf{x}_j)\big]. \quad (17)$$

$I(\mathbf{y};\mathbf{x}_k|\mathbf{x}_j)$ quantifies the amount of the classification information that $\mathbf{x}_k$ provides when $\mathbf{x}_j$ has been selected [40]. This information cannot be provided by $\mathbf{S}$. Compared with $I(\mathbf{y};\mathbf{x}_k)$, $I(\mathbf{y};\mathbf{x}_k|\mathbf{x}_j)$ does not involve the redundant information of pairwise features for classification.

Some methods, which aim to reduce redundancy, can be transformed into the methods that select features with large new classification information according to Eq. (9) [26]. When examining a candidate feature $\mathbf{x}_k$, increasing $I(\mathbf{y};\mathbf{x}_k|\mathbf{x}_j)$ is equivalent to decreasing $I(\mathbf{y};\mathbf{x}_j;\mathbf{x}_k)$. However, $I(\mathbf{y};\mathbf{x}_k|\mathbf{x}_j) > I(\mathbf{y};\mathbf{x}_l|\mathbf{x}_j)$ does not necessarily mean $I(\mathbf{y};\mathbf{x}_j;\mathbf{x}_k) < I(\mathbf{y};\mathbf{x}_j;\mathbf{x}_l)$ when two different candidate features $\mathbf{x}_k$ and $\mathbf{x}_l$ are evaluated. This finding implies that maximizing new classification information does not guarantee minimizing redundancy.

In light of the above analysis, *ICI* is introduced in the next section. *ICI* assembles redundancy information and new classification information into one term. Thus, both evaluation criteria play critical roles simultaneously in finding highly predictive as well as lowly redundant features.

3 Independent Classification Information

3.1 Definition

Definition 1 (Independent Classification Information). *Suppose features $\mathbf{x}_1$ and $\mathbf{x}_2$ are involved in recognizing the target class $\mathbf{y}$. Then, their independent classification information is defined as*

$$ICI(\mathbf{y};\mathbf{x}_1,\mathbf{x}_2) = I(\mathbf{y};\mathbf{x}_1|\mathbf{x}_2) + I(\mathbf{y};\mathbf{x}_2|\mathbf{x}_1). \quad (18)$$

ICI focuses on the amount of the specific classification information provided by a feature when another feature is given. Suppose one feature is a candidate and the other feature is selected, *ICI* indicates the amount of the new classification information provided by the candidate feature and the amount of the classification information preserved by the selected feature.

Mutual information between feature and class and between feature and feature should be further investigated to understand what is measured by *ICI*. Fig. 2 shows two cases of *ICI* of pairwise features, which are marked as shadowed parts. For case I, two features, namely, $\mathbf{x}_1$ and $\mathbf{x}_2$, are statistically independent from each other, i.e., $p(\mathbf{x}_1,\mathbf{x}_2) = p(\mathbf{x}_1)p(\mathbf{x}_2)$. Their classification information is not correlated with each other, i.e., $I(\mathbf{y};\mathbf{x}_1|\mathbf{x}_2) = I(\mathbf{y};\mathbf{x}_1)$ and $I(\mathbf{y};\mathbf{x}_2|\mathbf{x}_1) = I(\mathbf{y};\mathbf{x}_2)$. That is, their information for predicting classes is exactly the summation of their respective mutual information with classes. In this case, $ICI(\mathbf{y};\mathbf{x}_1,\mathbf{x}_2) = I(\mathbf{y};\mathbf{x}_1) + I(\mathbf{y};\mathbf{x}_2)$.

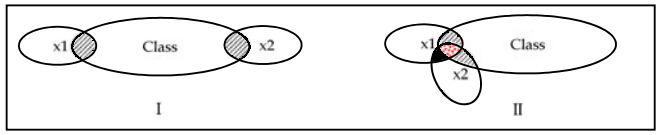


Fig. 2. Independent classification information offered by two features for two cases (marked as the shadowed parts): For case I, two features are statistically independent, and they are partially dependent for case II.

For case II, two features tightly or loosely correlate with each other, which is common in feature selection. The total classification information provided by two features can be separated to two parts, namely, *ICI* and dependent classification information. *ICI* represents the unshared information and comprises two terms, namely, $I(\mathbf{y};\mathbf{x}_1|\mathbf{x}_2)$ and $I(\mathbf{y};\mathbf{x}_2|\mathbf{x}_1)$. Each term represents the different predictive information of one feature from another feature. Hence, both terms provided respectively by each feature are distinct and helpful for recognizing the target class. They are asymmetric, and cannot be replaced by each other.

Another information is the dependent one, which is depicted as the red point part in Fig. 2. This information is the same as that shared by two features. From another angle, this information is the interaction of two features with the target class, which is exactly $I(\mathbf{y};\mathbf{x}_1;\mathbf{x}_2)$, i.e., the class-relevant redundancy provided by one feature if another feature is selected. In other words, this information fails to help enhance the predictive ability of a subset when a candidate feature is added.

The overlapped area in case II also includes a part unrelated to classification, which is exactly $I(\mathbf{x}_1;\mathbf{x}_2|\mathbf{y})$ and marked black in Fig. 2. This information is also a part of the relevance of two features, and is counted as feature redundancy by some selection methods. In fact, this part positively contributes to the joint predictive ability of two features, because large $I(\mathbf{x}_1;\mathbf{x}_2|\mathbf{y})$ means small $I(\mathbf{y};\mathbf{x}_1;\mathbf{x}_2)$. Therefore, directly employing the mutual information of two features as their redundancy cannot reflect their actual relationship in classification. One feature redundant with another feature fails to indicate that both features preserve little different classification information.

3.2 Combination of Feature Redundancy and New Classification Information

Feature redundancy and new classification information are investigated separately by two feature selection groups on the basis of mutual information. Both terms are essential for selecting predictive features that are lowly redundant or highly complementary to each other. Thus, the relations between two kinds of information and *ICI* are further investigated in this section.

Suppose $\mathbf{S} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{k-1}\}$ is selected. Then, *ICI* between the candidate feature $\mathbf{x}_k$ and $\mathbf{S}$ is calculated as

$$\sum_{\mathbf{x}_j \in \mathbf{S}} ICI(\mathbf{y};\mathbf{x}_j,\mathbf{x}_k) = \sum_{\mathbf{x}_j \in \mathbf{S}} \big[I(\mathbf{y};\mathbf{x}_j|\mathbf{x}_k) + I(\mathbf{y};\mathbf{x}_k|\mathbf{x}_j)\big]. \quad (19)$$

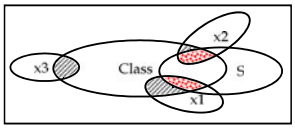


Fig. 3. Relationship between three features and subset S in recognizing the target class: $\mathbf{x}_1$ and $\mathbf{x}_2$ take equal class-relevant redundancy (red point parts) for S, and $\mathbf{x}_1$ and $\mathbf{x}_3$ provide equal new information (shadowed parts) for S.

Clearly, the term $\sum_{\mathbf{x}_j \in \mathbf{S}} I(\mathbf{y}; \mathbf{x}_k | \mathbf{x}_j)$ corresponds to the new classification information provided by $\mathbf{x}_k$ for $\mathbf{S}$, which is measured in the new classification information-based methods as mentioned in Section 2.

As to the term $I(\mathbf{y}; \mathbf{x}_j | \mathbf{x}_k)$, it holds that $I(\mathbf{y}; \mathbf{x}_j | \mathbf{x}_k) + I(\mathbf{y}; \mathbf{x}_j; \mathbf{x}_k) = I(\mathbf{y}; \mathbf{x}_j)$. In feature selection, $I(\mathbf{y}; \mathbf{x}_j)$ can be regarded as constant. Thus, $I(\mathbf{y}; \mathbf{x}_j | \mathbf{x}_k)$ is negatively correlated with $I(\mathbf{y}; \mathbf{x}_j; \mathbf{x}_k)$. It implies that maximizing $I(\mathbf{y}; \mathbf{x}_j | \mathbf{x}_k)$ is equivalent to minimizing $I(\mathbf{y}; \mathbf{x}_j; \mathbf{x}_k)$. Therefore, the term $\sum_{\mathbf{x}_j \in \mathbf{S}} I(\mathbf{y}; \mathbf{x}_j | \mathbf{x}_k)$ in *ICI* corresponds to the feature redundancy $\sum_{\mathbf{x}_j \in \mathbf{S}} I(\mathbf{y}; \mathbf{x}_j; \mathbf{x}_k)$ taken by $\mathbf{x}_k$ for $\mathbf{S}$.

In summary, $\sum_{\mathbf{x}_j \in \mathbf{S}} ICI(\mathbf{y}; \mathbf{x}_j, \mathbf{x}_k)$ focuses on both new classification information and redundant information, which are separately considered by common mutual information-based methods. *ICI* prefers features that are different from each other in classification.

To further demonstrate the necessity of considering both information terms in *ICI*, the example in Fig. 1 is extended to illustrate two cases in feature selection, as shown in Fig. 3. Features $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$ are evaluated in terms of their classification correlations with $\mathbf{S}$. Suppose $\mathbf{x}_1$ and $\mathbf{x}_2$ take equal class-relevant redundancy for $\mathbf{S}$, i.e., $I(\mathbf{y}; \mathbf{S}; \mathbf{x}_1) = I(\mathbf{y}; \mathbf{S}; \mathbf{x}_2)$, which is represented as two red point parts with equal area. $\mathbf{x}_1$ and $\mathbf{x}_3$ provide equal new classification information for $\mathbf{S}$, i.e., $I(\mathbf{y}; \mathbf{x}_1 | \mathbf{S}) = I(\mathbf{y}; \mathbf{x}_3 | \mathbf{S})$, which is represented as equal areas of two shadowed parts. For $\mathbf{x}_1$ and $\mathbf{x}_2$, it holds that $I(\mathbf{y}; \mathbf{S}) - I(\mathbf{y}; \mathbf{S} | \mathbf{x}_1) = I(\mathbf{y}; \mathbf{S}) - I(\mathbf{y}; \mathbf{S} | \mathbf{x}_2)$. This implies that $\mathbf{S}$ preserves the same classification information whichever feature is selected. As demonstrated in Fig. 3, $\mathbf{x}_1$ is more related to classes than $\mathbf{x}_2$. This finding suggests that the total information of $\mathbf{S}$ is more increased when $\mathbf{x}_1$ is added than when $\mathbf{x}_2$ is added. Therefore, *ICI* prefers $\mathbf{x}_1$, which is more relevant to classes, rather than $\mathbf{x}_2$.

For features $\mathbf{x}_1$ and $\mathbf{x}_3$, $I(\mathbf{y}; \mathbf{x}_1 | \mathbf{S}) = I(\mathbf{y}; \mathbf{x}_3 | \mathbf{S})$ indicates that $\mathbf{x}_1$ and $\mathbf{x}_3$ provide the same new classification information for $\mathbf{S}$ when either of them is selected. An obvious case is that $I(\mathbf{S}; \mathbf{x}_1) > I(\mathbf{S}; \mathbf{x}_3) = 0$. It can be interpreted as $\mathbf{x}_3$ statistically independent from any feature in $\mathbf{S}$. Or in other words, $\mathbf{x}_1$ is more relevant to $\mathbf{S}$ than $\mathbf{x}_3$. This directly leads to $I(\mathbf{y}; \mathbf{S} | \mathbf{x}_3) > I(\mathbf{y}; \mathbf{S} | \mathbf{x}_1)$, which indicates that the unshared classification information of $\mathbf{S}$ is more preserved if $\mathbf{x}_3$ is selected. Therefore, *ICI* prefers $\mathbf{x}_3$ rather than $\mathbf{x}_1$.

Choosing $\mathbf{x}_1$ or $\mathbf{x}_3$ is a dilemma for the new classification information-based selection methods, because both features provide the same new information. Feature $\mathbf{x}_1$ provides more classification information than $\mathbf{x}_3$, i.e., $I(\mathbf{y}; \mathbf{x}_1) > I(\mathbf{y}; \mathbf{x}_3)$. Meanwhile, $\mathbf{x}_1$ is more redundant to $\mathbf{S}$ than $\mathbf{x}_3$. Therefore, this group of methods has to face the problem of choosing high relevance or low redundancy. In contrast, *ICI* definitely chooses $\mathbf{x}_3$.

3.3 A Loose Upper Bound

As aforementioned, the aim of feature selection is to select a set of features $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ that can maximize $I(\mathbf{y}; \mathbf{x}_1, \ldots, \mathbf{x}_k)$ in Eq. (2). To show how maximizing *ICI* is beneficial to this aim, Theorem 1 is given as follows:

Theorem 1. *Suppose* $\mathbf{S} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{k-1}\}$ *is selected. Let* $\widetilde{\mathbf{S}} = \mathbf{S} \bigcup \{\mathbf{x}_k\}$, *and* $\widehat{\mathbf{S}}_j = \widetilde{\mathbf{S}} \setminus \{\mathbf{x}_j\}$. *Then,* $\sum_{\mathbf{x}_j \in \mathbf{S}} ICI(\mathbf{y}; \mathbf{x}_j, \mathbf{x}_k)$ *is a loose upper bound of* $I(\mathbf{y}; \widetilde{\mathbf{S}})$.

Proof. According to the mutual information chain rule, $I(\mathbf{y}; \widetilde{\mathbf{S}}) = I(\mathbf{y}; \mathbf{x}_j | \widehat{\mathbf{S}}_j\}) + I(\mathbf{y}; \widehat{\mathbf{S}}_j\}), \forall \mathbf{x}_j \in \widetilde{\mathbf{S}}$. Then, $I(\mathbf{y}; \widetilde{\mathbf{S}}) = \frac{1}{k} \sum_{\mathbf{x}_j \in \widetilde{\mathbf{S}}} [I(\mathbf{y}; \mathbf{x}_j | \widehat{\mathbf{S}}_j) + I(\mathbf{y}; \widehat{\mathbf{S}}_j)] = \frac{1}{k} \sum_{\mathbf{x}_j \in \widetilde{\mathbf{S}}} [H(\mathbf{y} | \widehat{\mathbf{S}}_j) - H(\mathbf{y} | \widetilde{\mathbf{S}}) + H(\mathbf{y}) - H(\mathbf{y} | \widehat{\mathbf{S}}_j)]$.

Because $H(\mathbf{y} | \widetilde{\mathbf{S}}) \leq H(\mathbf{y} | \widehat{\mathbf{S}}_j) \leq [H(\mathbf{y} | \mathbf{x}_i)_{j \neq i} + H(\mathbf{y} | \mathbf{x}_k)]$, it is obtained that $I(\mathbf{y}; \widetilde{\mathbf{S}}) \leq \frac{1}{k} \sum_{\mathbf{x}_i \in \widetilde{\mathbf{S}}} [H(\mathbf{y} | \mathbf{x}_i) + H(\mathbf{y} | \mathbf{x}_k) - 2H(\mathbf{y} | \widetilde{\mathbf{S}}) + H(\mathbf{y})]$. Because $H(\mathbf{y} | \widetilde{\mathbf{S}}) = H(\mathbf{y} | \mathbf{x}_i, \mathbf{x}_k) - I(\mathbf{y}; \widehat{\mathbf{S}}_{i,k} | \mathbf{x}_i, \mathbf{x}_k) \geq H(\mathbf{y} | \mathbf{x}_i, \mathbf{x}_k) + H(\mathbf{x}_i, \mathbf{x}_k) - H(\widetilde{\mathbf{S}}) \geq H(\mathbf{y} | \mathbf{x}_i, \mathbf{x}_k) - H(\widehat{\mathbf{S}}_{i,k})$, it holds that $I(\mathbf{y}; \widetilde{\mathbf{S}}) \leq \frac{1}{k} \sum_{\mathbf{x}_i \in \widetilde{\mathbf{S}}} [I(\mathbf{y}; \mathbf{x}_i | \mathbf{x}_k) + I(\mathbf{y}; \mathbf{x}_k | \mathbf{x}_i) + H(\mathbf{y}) + 2H(\widehat{\mathbf{S}}_{i,k})] = \frac{1}{k} [\sum_{\mathbf{x}_j \in \mathbf{S}} ICI(\mathbf{y}; \mathbf{x}_j, \mathbf{x}_k) + 2H(\mathbf{S}) + kH(\mathbf{y}) + 2 \sum_{\mathbf{x}_j \in \mathbf{S}} H(\mathbf{S} \setminus \{\mathbf{x}_j\})]$. Clearly, $\sum_{\mathbf{x}_j \in \mathbf{S}} H(\mathbf{S} \setminus \{\mathbf{x}_j\})$, $H(\mathbf{S})$, and $H(\mathbf{y})$ can be considered constant in the selection process. Therefore, $\sum_{\mathbf{x}_j \in \mathbf{S}} ICI(\mathbf{y}; \mathbf{x}_j, \mathbf{x}_k)$ is a loose upper bound of $I(\mathbf{y}; \widetilde{\mathbf{S}})$. □

According to Theorem 1, *ICI* of the candidate feature with the selected subset is an upper bound of the total classification information of feature subset. Therefore, maximizing *ICI* may help enhance the global classification performance of the subset.

ICI focuses on the unshared mutual information of features with classes, which is exactly the global discriminative information provided by features. When a candidate feature is evaluated, its particular classification information and the specific information preserved by the selected features should be prior considered. The classification information of the candidate feature that has already given by the selected features is redundant. Thus, the global predictive performance of the feature subset is expected to be enhanced by maximizing its *ICI*. This strategy is suggested as reasonable by the theoretical analysis of Theorem 1.

4 Max-Relevance and Max-Independence

4.1 Definition and Property

Suppose $\mathbf{S} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{k-1}\}$ is selected. The criterion of *Max-Relevance and Max-Independence* (*MRI*) is presented as

$$J_{MRI}(\mathbf{x}_k) = I(\mathbf{y}; \mathbf{x}_k) + \sum_{\mathbf{x}_j \in \mathbf{S}} ICI(\mathbf{y}; \mathbf{x}_j, \mathbf{x}_k). \tag{20}$$

MRI considers both feature relevance with classes and feature classification independence between each other. A candidate feature with high *MRI* score tends to exhibit an excellent discriminative ability that is different from those provided by the selected features. Some properties of *MRI* are discussed as follows.

Proposition 1. *Suppose* $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ *are i.i.d.. Then, MRI is equivalent to the Max-Relevance criterion.*

Proof. According to the assumption of i.i.d. variables, it holds that $I(\mathbf{y};\mathbf{x}_i|\mathbf{x}_j) = I(\mathbf{y};\mathbf{x}_i), \forall i,j = 1,\ldots,k, i \neq j$. In this case, *MRI* can be reformulated as $J_{MRI}(\mathbf{x}_k) = I(\mathbf{y};\mathbf{x}_k) + \sum_{\mathbf{x}_j \in \mathbf{S}} \big[I(\mathbf{y};\mathbf{x}_j) + I(\mathbf{y};\mathbf{x}_k)\big] = kI(\mathbf{y};\mathbf{x}_k) + \sum_{\mathbf{x}_j \in \mathbf{S}} I(\mathbf{y};\mathbf{x}_j)$. In feature selection process, $\sum_{\mathbf{x}_j \in \mathbf{S}} I(\mathbf{y};\mathbf{x}_j)$ can be considered constant. Then, it holds that $J_{MRI}(\mathbf{x}_k) \propto kI(\mathbf{y};\mathbf{x}_k) \propto I(\mathbf{y};\mathbf{x}_k)$. Hence, *MRI* is equivalent to the Max-relevance criterion. □

Proposition 1 claims a special case that *MRI* is equivalent to the *Max-relevance* criterion, that is, features are statistically independent from each other. In this case, features independently provide classification information that cannot be offered by other features. *MRI* prefers the most relevant features in this case.

Proposition 2. *Suppose* $H(\mathbf{y}|\mathbf{x}_i,\mathbf{x}_j) = H(\mathbf{y}|\mathbf{x}_i), \forall i,j = 1,\ldots,k$ *and* $i \leq j$. *Then, MRI is equivalent to* $-\sum_{\mathbf{x}_j \in \mathbf{S}} I(\mathbf{y};\mathbf{x}_j;\mathbf{x}_k)$ *when* $k \geq 3$.

Proof. According to the assumption of $H(\mathbf{y}|\mathbf{x}_i,\mathbf{x}_j) = H(\mathbf{y}|\mathbf{x}_i), \forall i,j=1,\ldots,k$ and $i \leq j$, it holds that $I(\mathbf{y};\mathbf{x}_k|\mathbf{x}_j) = H(\mathbf{y}|\mathbf{x}_j) - H(\mathbf{y}|\mathbf{x}_j,\mathbf{x}_k) = 0,\ \forall \mathbf{x}_j \in \mathbf{S}$. Then, it is obtained that $\sum_{\mathbf{x}_j \in \mathbf{S}} I(\mathbf{y};\mathbf{x}_j;\mathbf{x}_k) = \sum_{\mathbf{x}_j \in \mathbf{S}} [I(\mathbf{y};\mathbf{x}_k|\mathbf{x}_j) - I(\mathbf{y};\mathbf{x}_k)] = -(k-1)I(\mathbf{y};\mathbf{x}_k)$. Thus, it holds for MRI that $J_{MRI}(\mathbf{x}_k) = I(\mathbf{y};\mathbf{x}_k) + \sum_{\mathbf{x}_j \in \mathbf{S}} I(\mathbf{y};\mathbf{x}_j|\mathbf{x}_k) = \sum_{\mathbf{x}_j \in \mathbf{S}} I(\mathbf{y};\mathbf{x}_j,\mathbf{x}_k) - (k-2)I(\mathbf{y};\mathbf{x}_k) = \sum_{\mathbf{x}_j \in \mathbf{S}} I(\mathbf{y};\mathbf{x}_j) - (k-2)\ I(\mathbf{y};\mathbf{x}_k)$. In feature selection process, $\sum_{\mathbf{x}_j \in \mathbf{S}} I(\mathbf{y};\mathbf{x}_j)$ can be considered constant. Hence, $J_{MRI}(\mathbf{x}_k) \propto -(k-2)I(\mathbf{y};\mathbf{x}_k) \propto -I(\mathbf{y};\mathbf{x}_k) \propto -\sum_{\mathbf{x}_j \in \mathbf{S}} I(\mathbf{y};\mathbf{x}_j;\mathbf{x}_k)$, when $k \geq 3$. □

Proposition 2 claims another special case, i.e., no new classification information can be provided by the newly added features, which is similar to the Example 2 in Section 1. This case can be regarded as an approximate Markov blanket with information measure [9], [41]. That is, the previously selected features subsume the classification information that the later selected features provide. In this case, *MRI* is equivalent to minimizing multi-information between candidate features and selected features. As aforementioned, less multi-information means less redundancy. Thus, *MRI* prefers the least redundant features in this case.

Proposition 3. *Suppose* $\widetilde{\mathbf{S}} = \mathbf{S} \bigcup \{\mathbf{x}_k\}$. *Then, MRI conduces to maximizing* $I(\mathbf{y};\widetilde{\mathbf{S}})$.

Proof. According to the monotonicity of conditional entropy, $H(\mathbf{y}|\widetilde{\mathbf{S}}) \leq H(\mathbf{y}|\mathbf{x}_j), \forall \mathbf{x}_j \in \widetilde{\mathbf{S}}$. Then, it is obtained that $H(\mathbf{y}) - H(\mathbf{y}|\widetilde{\mathbf{S}}) \geq H(\mathbf{y}) - H(\mathbf{y}|\mathbf{x}_j)$. That is, $I(\mathbf{y};\widetilde{\mathbf{S}}) \geq I(\mathbf{y};\mathbf{x}_j)$. Hence, $I(\mathbf{y};\widetilde{\mathbf{S}}) \geq \frac{1}{k}\sum_{j=1}^{k} I(\mathbf{y};\mathbf{x}_j)$. Therefore, $\sum_{\mathbf{x}_j \in \widetilde{\mathbf{S}}} I(\mathbf{y};\mathbf{x}_j)$ is the lower bound of $I(\mathbf{y};\widetilde{\mathbf{S}})$. In feature selection process, $\sum_{\mathbf{x}_j \in \mathbf{S}} I(\mathbf{y};\mathbf{x}_j)$ can be considered constant. Thus, maximizing $I(\mathbf{y};\mathbf{x}_k)$ is equivalent to maximizing the lower bound of $I(\mathbf{y};\widetilde{\mathbf{S}})$. According to Theorem 1, $\sum_{\mathbf{x}_j \in \mathbf{S}} ICI(\mathbf{y};\mathbf{x}_j,\mathbf{x}_k)$ is a loose upper bound of $I(\mathbf{y};\widetilde{\mathbf{S}})$. Hence, *MRI* conduces to maximizing $I(\mathbf{y};\widetilde{\mathbf{S}})$ by simultaneously maximizing its lower and upper bounds. □

Proposition 3 claims that *MRI* helps maximize the total classification information of feature subset. According to the Fano's equality [42], the lower and upper bounds of Bayesian classification error Pe are both correlated with the conditional entropy of class when $\widetilde{\mathbf{S}}$ is considered as a condition, which is given as $\frac{H(\mathbf{y}|\widetilde{\mathbf{S}})-1}{\log|\mathbf{y}|} \leq Pe \leq \frac{H(\mathbf{y}|\widetilde{\mathbf{S}})}{2}$ [43], [44]. Because *MRI* conduces to maximizing $I(\mathbf{y};\widetilde{\mathbf{S}})$ as demonstrated in Proposition 3, and $H(\mathbf{y}|\widetilde{\mathbf{S}}) = H(\mathbf{y}) - I(\mathbf{y};\widetilde{\mathbf{S}})$, *MRI* conduces to minimizing Bayesian classification error. Proposition 3 declaims that *MRI* is expected to select the highly discriminative feature subset.

In terms of the computation of *MRI*, it mainly lies on estimating $ICI(\mathbf{y};\mathbf{x}_j,\mathbf{x}_k)$ that involves two terms as follows:

$$I(\mathbf{y};\mathbf{x}_j|\mathbf{x}_k) = \sum_{\mathbf{y}}\sum_{\mathbf{x}_j}\sum_{\mathbf{x}_k} p(\mathbf{y}|\mathbf{x}_j,\mathbf{x}_k)p(\mathbf{x}_j,\mathbf{x}_k)\log\frac{p(\mathbf{y}|\mathbf{x}_j,\mathbf{x}_k)}{p(\mathbf{y}|\mathbf{x}_k)},$$

$$I(\mathbf{y};\mathbf{x}_k|\mathbf{x}_j) = \sum_{\mathbf{y}}\sum_{\mathbf{x}_k}\sum_{\mathbf{x}_j} p(\mathbf{y}|\mathbf{x}_j,\mathbf{x}_k)p(\mathbf{x}_j,\mathbf{x}_k)\log\frac{p(\mathbf{y}|\mathbf{x}_j,\mathbf{x}_k)}{p(\mathbf{y}|\mathbf{x}_j)}.$$

Apparently, only one item is different when comparing two equations, i.e., conditional class probability provided by each feature. The other probability items are the same, which facilitates efficient computation of *MRI*.

4.2 Relationship Between *MRI* and Some Popular Mutual Information-Based Evaluation Criteria

Some mutual information-based methods can be regarded as a specific form of the framework described in Eq. (10), when relevance and redundancy are considered at different degrees to evaluate features. Hence, *MRI* should be compared with some typical criteria to observe their differences in feature evaluation.

For comparison, *MRI* is reformulated as follows:

$$J_{MRI}(\mathbf{x}_k) = I(\mathbf{y};\mathbf{x}_k) + \sum_{\mathbf{x}_j \in \mathbf{S}} \big[I(\mathbf{y};\mathbf{x}_k) - 2I(\mathbf{y};\mathbf{x}_j;\mathbf{x}_k) + I(\mathbf{y};\mathbf{x}_j)\big] = kI(\mathbf{y};\mathbf{x}_k) - 2\sum_{\mathbf{x}_j \in \mathbf{S}} I(\mathbf{y};\mathbf{x}_j;\mathbf{x}_k) + \sum_{\mathbf{x}_j \in \mathbf{S}} I(\mathbf{y};\mathbf{x}_j).$$

Note that $\sum_{\mathbf{x}_j \in \mathbf{S}} I(\mathbf{y};\mathbf{x}_j)$ can be considered constant in feature selection process. Thus, it holds for *MRI* that

$$J_{MRI}(\mathbf{x}_k) \propto I(\mathbf{y};\mathbf{x}_k) - \frac{2}{|\mathbf{S}|+1}\sum_{\mathbf{x}_j \in \mathbf{S}} I(\mathbf{y};\mathbf{x}_j;\mathbf{x}_k). \tag{21}$$

MRI can reduce feature redundancy in selection process. In contrast to the redundancy reduction-based criteria, such as *CIFE* in Eq. (14) and *ICAP* in Eq. (15), *MRI* varies its balance weight between the relevance term and the redundancy term when a feature subset is expanded. This weight is negatively correlated with the cardinality of feature subset. A similar strategy is adopted by some other evaluation criteria, such as *mRMR* and *mIMR*. They average the redundancy term to pursue an equal importance between feature relevance and feature redundancy. Therefore, the effect of redundancy in feature evaluation does not greatly exceed relevance when a large number of features are added to subset. This mechanism helps prevent the selection of lowly redundant but also weakly relevant features.

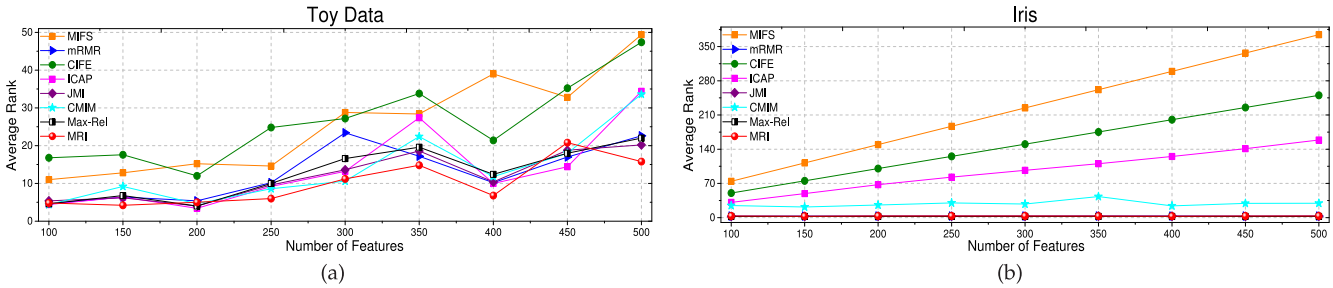


Fig. 4. Average ranks of the relevant features selected by each criterion (the lower of the ranks, the better).

From another aspect, *MRI* can also be reformulated as

$$J_{MRI}(\mathbf{x}_k) = I(\mathbf{y};\mathbf{x}_k) + \sum_{\mathbf{x}_j \in \mathbf{S}} \big[2I(\mathbf{y};\mathbf{x}_j,\mathbf{x}_k) - I(\mathbf{y};\mathbf{x}_k) - I(\mathbf{y};\mathbf{x}_j)\big] \propto I(\mathbf{y};\mathbf{x}_k) + \sum_{\mathbf{x}_j \in \mathbf{S}} \big[2I(\mathbf{y};\mathbf{x}_j,\mathbf{x}_k) - I(\mathbf{y};\mathbf{x}_k)\big].$$

Suppose $k \gg 1$. Then, it holds for *MRI* that

$$J_{MRI}(\mathbf{x}_k) \propto \sum_{\mathbf{x}_j \in \mathbf{S}} I(\mathbf{y};\mathbf{x}_k,\mathbf{x}_j) + \sum_{\mathbf{x}_j \in \mathbf{S}} I(\mathbf{y};\mathbf{x}_j|\mathbf{x}_k). \tag{22}$$

In contrast to the new classification information-based criteria, such as *JMI* in Eq. (16), *MRI* additionally considers the preserved information of the selected feature subset. *MRI* prefers features that can provide maximal new information and meanwhile help the selected features preserve maximal unshared classification information. Therefore, *MRI* pursues features that are independent in recognizing classes.

In fact, within the framework of Eq. (10), *JMI* can also be rewritten as

$$J_{JMI}(\mathbf{x}_k) \propto I(\mathbf{y};\mathbf{x}_k) - \frac{1}{|\mathbf{S}|}\sum_{\mathbf{x}_j \in \mathbf{S}} I(\mathbf{y};\mathbf{x}_j;\mathbf{x}_k). \tag{23}$$

JMI employs the same strategy as that in $mRMR$ to regulate the effects of relevance and redundancy in feature evaluation. The difference between *JMI* and *mRMR* is that $I(\mathbf{y};\mathbf{x}_j;\mathbf{x}_k)$ and $I(\mathbf{x}_j;\mathbf{x}_k)$ are considered as redundancy by two criteria, respectively. Note that *JMI* is close to *MRI* to some extent, which indicates that they may show some similarities in feature selection.

The above comparisons reveal that the effects of relevancy and redundancy are regulated differently when parameters λ and γ of *CMI* in Eq. (10) vary. However, the optimal values of these parameters are hard to determine. *MRI* regulates the effects of relevance and redundancy through a mechanism different from other mutual information-based criteria, i.e., maximizing the new classification information provided by a candidate feature and the information preserved by the selected features simultaneously.

5 Experiment and Analysis

This section mainly compares *MRI* with several well-known mutual information-based feature evaluation criteria, i.e., *MIFS* [11], *mRMR* [12], *CIFE* [14], *ICAP* [17], *JMI* [22], *CMIM* [25] and *Max-Relevance*. The parameter β of *MIFS* is set to 1 according to the suggestion of the author. An efficient greedy searching strategy, namely, Sequential Forward Search (SFS), is employed for all the selection criteria for a fair comparison. SFS starts from an empty feature subset, and includes the most relevant feature at first. Then, it expands subset with one excellent feature highly scored by evaluation criterion at each selection step. In terms of the feature searching complexity, it is approximately equal to $O(kd)$ for the compared criteria, where k and d are the numbers of the selected and original features, respectively.

First, the capabilities of the compared criteria to find significant features from noisy features are tested on the artificial and real-world data sets. Meanwhile, the effective information provided by the top selected features is counted. Second, noisy instances are added to a real-world data set Gas, and the selection performance of the feature evaluation criteria is compared according to diverse metrics. Third, 11 groups of UCI data sets and 11 groups of Microarray expression data sets are employed to test the classification performance of feature subsets, and their stability and inconsistency indices are also assessed. Fourth, parameter sensitivity tests are conducted between *MIFS* and *MRI*. Last, several popular non-mutual-information-based feature selection approaches are compared with *MRI*.

5.1 Exp. 1: Varying the Size of Noisy Features

Several groups of toy data sets that are randomly constructed in the spider[1] environment and Iris data set fetched from UCI [27] are used to test the performance of the compared criteria in selecting target features from noisy features. Only the first five features in toy data and the original four features in Iris are relevant to classes, and the other ones are noisy features, which are weakly relevant or irrelevant to classes. The number of features increases from 100 to 500 in the interval of 50.

First, the compared criteria sort features and record the average ranks of the five/four relevant features on the ranking lists. A small value corresponds to high ranks of the relevant features, which indicates that the corresponding criterion is effective in selecting target features. Conversely, a large value reflects the poor ability in tackling noisy features. Experimental results are shown in Fig. 4.

The relevant features tend to be lowly scored when more noisy features are added in Fig. 4. In this case, *MRI* assigns relevant features with relatively smaller observation values than the other criteria on toy data, and also performs well and comparably stable as *mRMR*, *JMI*, and *Max-Relevance* on Iris data. *MIFS*, *CIFE*, and *ICAP* show inferior to the other compared criteria. The lowly redundant but also weakly relevant features are probably selected by these criteria, and the relevant features are lowly preferred.

1. http://people.kyb.tuebingen.mpg.de/spider/main.html.

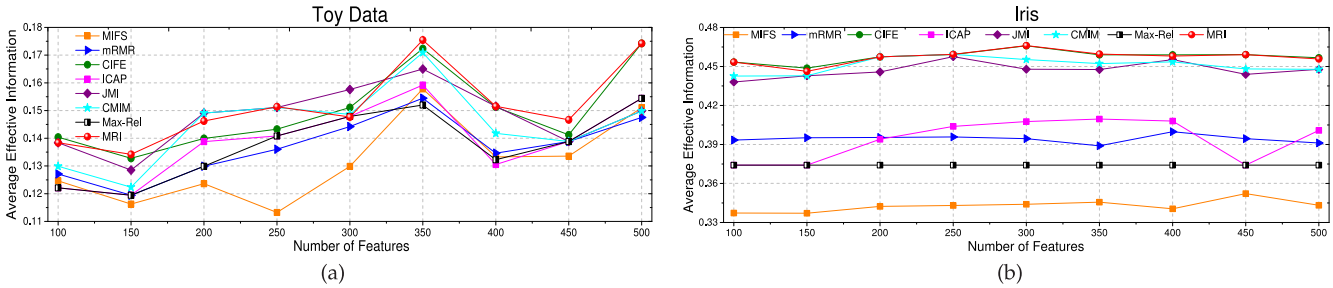


Fig. 5. Average effective information of the top ranked features (the higher of the effective information, the better).

Second, the top-5 ranked features on toy data and top-4 features on Iris data are compared in terms of the average effective classification information, defined as

$$ECI(\mathbf{S}) = \frac{1}{|\mathbf{S}|}\sum_{\mathbf{x}_i \in \mathbf{S}} I(\mathbf{y};\mathbf{x}_i) - \frac{2}{|\mathbf{S}|(|\mathbf{S}|-1)}\sum_{\mathbf{x}_i,\mathbf{x}_j \in \mathbf{S}} I(\mathbf{y};\mathbf{x}_i;\mathbf{x}_j). \quad (24)$$

The former term is the average feature relevance with target class, and the latter term is the average redundancy between features. Clearly, informative features provide high ECIs. The experimental results are depicted in Fig. 5.

ECI measures the effective classification information of selected features. The weakly relevant but lowly redundant features may provide equal ECIs with those strongly relevant but highly redundant features. Thus, the redundancy reduction-based criteria obtain relatively high ECIs, such as *CIFE*. The new information-based criteria also perform well due to their selections of relevant features providing unshared information, such as *JMI* and *CMIM*. Note that the performance of *Max-Relevance* in the ECI test is worse than its performance in the ranking test. The strategy of merely selecting the most relevant features weakens the performance of *Max-Relevance* in reducing redundancy. In contrast, *MRI* prefers strongly relevant as well as lowly redundant features. As a result, *MRI* yields excellent results in the ECI test.

5.2 Exp. 2: Varying the Size of Noisy Instances

A benchmark data set fetched from UCI, namely, Gas Sensor Array Drift Dataset, is used to test the performance of the compared criteria in selecting features on the data with noisy instances. A total of 13,910 instances are characterized by 128 features in Gas, and six pure gas classes are discriminated. 30 and 50 percent noisy instances are added to the original data in the weka environment [45], respectively. Each criterion selects k optimal features from three data sets, i.e., original data, 30 percent noisy data and 50 percent noisy data. k increases from 5 to 50 in the interval of 5. Three metrics, i.e., Representation Entropy [46], Inconsistency Rate [47], and Fisher Score [48], are exploited to measure the performance of feature subsets in terms of their discriminative information, consistency with the original data, and class separability, respectively.

Experimental results are demonstrated in Figs. 6, 7, and 8. The correlations between feature and class and between feature and feature change when noisy instances are added, which directly affects feature evaluations. That is, the optimal features in the original data are commonly different from the optimal features in the noisy data. Thus, the performance of feature subset selected by an evaluation criterion may vary when noisy instances are added. Fisher scores of the compared criteria tend to decline when 30 percent noisy instances are added to the original data. This indicates that the discriminative abilities of feature subsets are weakened on the noisy data. In this case, the declining speed of *MRI* is slower than those of the other criteria, and it is comparably good as *mRMR*. Nevertheless, *MRI* outperforms *mRMR* and the other criteria when noisy instances reach 50 percent.

In terms of inconsistency rate, it tends to increase when noisy instances are added. The average inconsistency rate of *MIFS* is 2.16×10^3 on the original data, and it increases to 2.59×10^3 on the 50 percent noisy data. The scores of inconsistency rate across three data sets actually increase, although the variation trend seems not to change obviously in Figs. 7a, 7b, and 7c. It demonstrates that the selected subsets are less consistent when noisy instances are added. In this case, the rising speeds of *MRI*, *ICAP*, and *CIFE* are slower than the other criteria.

Note that *JMI* and *MRI* illustrate their similarities in Figs. 6 and 8, i.e., their variation trends and metric scores are more similar to each other than the other criteria. In Fig. 7, *JMI* and *MRI* perform differently under the metric of inconsistency rate to some extent. Inconsistency rate can assess feature redundancy in the selected subset, which is not considered by representation entropy and Fisher score. Therefore, as demonstrated in Fig. 7, *MRI*, *ICAP*, and *CIFE* that are capable of reducing feature redundancy exhibit excellent under the metric of inconsistency rate. *JMI* is comparably inferior due to its attention on new classification information rather than feature redundancy.

We observe that a criterion tends to perform well under one metric whereas maybe bad under the other metrics. For instance, *CIFE* and *ICAP* are superior in terms of inconsistency rate, but inferior on Fisher score. Overall speaking, *MRI* exhibits promising results under different metrics. The performance of *MRI* will be further testified under some classification metrics in the next section.

5.3 Exp. 3: Comparing Classification Performance

Experiments on UCI data and Microarray expression data [27], [34], [49], [50] in Table 1 are conducted to test the classification performance of selected features. The benchmark data sets cover both binary-class and multi-class, and the number of original features varies from less than 50 to near to 50,000. The number of selected features, i.e., k, sequentially increases from 1 to 50 in the interval of 1. That is, the compared criteria respectively select 50 groups of feature subsets whose sizes increase from 1 to 50 for comparison. Note that the numbers of original features for the first five data sets in Table 1 are less than 50. In this case, k reaches up to the number of original features.

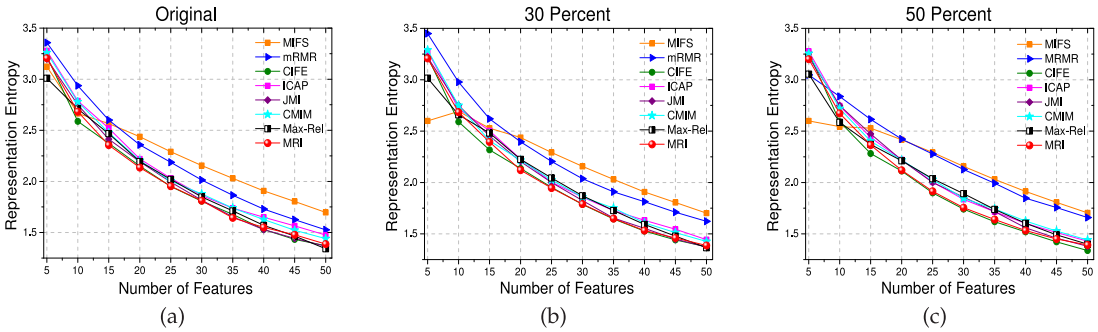


Fig. 6. Representation entropy of the selected feature subsets (the lower of the representation entropy, the better).

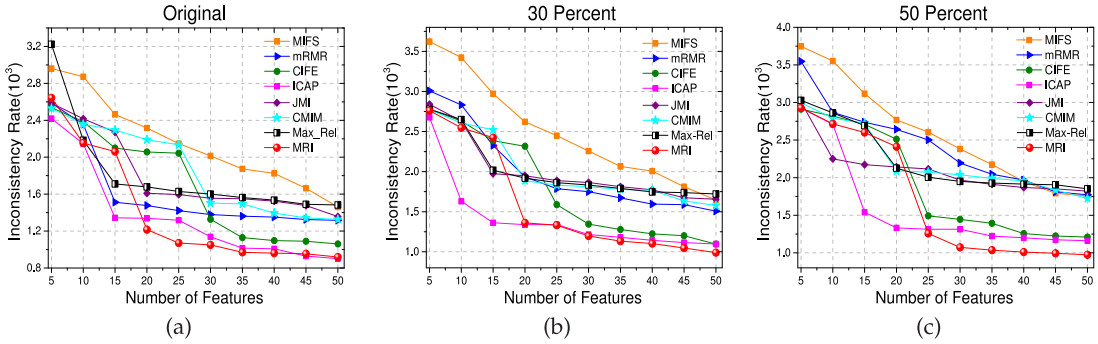


Fig. 7. Inconsistency rate of the selected feature subsets (the lower of the inconsistency rate, the better).

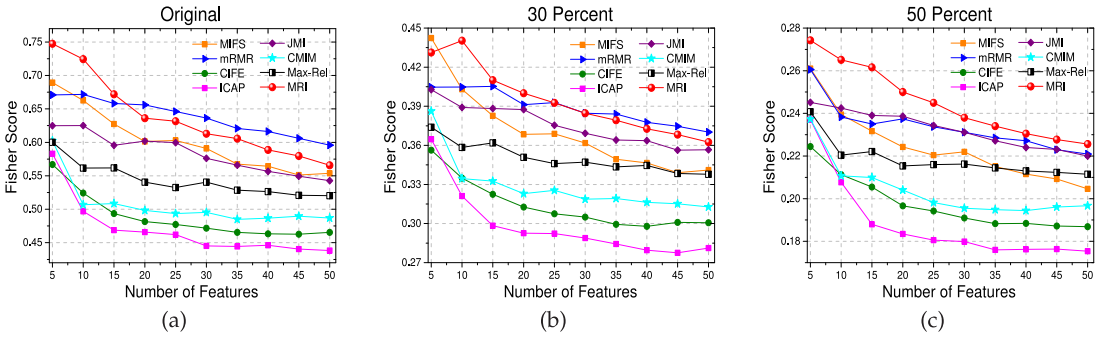


Fig. 8. Fisher score of the selected feature subsets (the higher of the Fisher score, the better).

Two classifiers are constructed by the selected features in the weka environment [45], i.e., 1-Nearest Neighbor (1-NN) classifier and Support Vector Machine (SVM) classifier, and tested with 10-fold cross-validations. Average classification accuracies of both classifiers across the 50 groups of feature subsets selected by each criterion are recorded in Tables 2 and 3. Furthermore, pairwise t-test at 5 percent significance level is conducted to evaluate the statistical significance of the results, and the best one and those not significantly worse than it in each row are highlighted in bold. Last two rows in each table record the average accuracies across all the benchmark data sets and the times of achieving best results for each evaluation criteria.

Note that all the compared criteria adopt filter selection strategies, which exclude induction algorithms in selection process. Hence, the performance of the compared criteria is unrelated to the choice of classifiers. In terms of the general performance, *MIFS* and *ICAP* perform better on the SVM classifier than on the 1-NN classier, while the other criteria show the opposite performance. However, this phenomenon is not general

TABLE 1
Benchmark Data Sets

Data set	#Features	#Instances	#Classes	Source
Image Segmentation	19	2,310	7	UCI
Phishing Websites	30	11,055	2	UCI
Ionosphere	34	351	2	UCI
Waveform	40	5,000	3	UCI
Connect-4	42	67,557	3	UCI
Nomao	120	34,465	2	UCI
Musk (Version1)	168	476	2	UCI
Lung	325	73	7	Microarray
UJIIndoorLoc	528	21,048	3	UCI
Smartphone Recognition	561	10,929	12	UCI
Internet Advertisements	1,558	3,279	2	UCI
Colon	2,000	62	2	Microarray
SRBCT	2,308	88	5	Microarray
DLBCL	4,026	88	6	Microarray
TOX-171	5,748	171	4	Microarray
Prostate_GE	5,966	102	2	Microarray
Breast	9,216	84	5	Microarray
Arcene	10,000	100	2	UCI
Cancers	12,533	174	11	Microarray
Leukemia	12,582	72	3	Microarray
GLI-85	22,283	85	2	Microarray
GLA-BRA-180	49,151	180	4	Microarray

TABLE 2
Average 1-NN Classification Accuracy (Mean±Std.) with p-Value (in Percentage)

Data	Algorithm							
	MIFS	*mRMR*	*CIFE*	*ICAP*	*JMI*	*CMIM*	*Max-Rel*	*MRI*
Image	90.14±7.57(0.00)	92.28±7.81(0.00)	93.73±7.91(0.00)	93.85±7.92(0.01)	**94.18±8.06(0.24)**	**94.19±8.05(0.14)**	92.75±8.79(0.02)	**94.23±8.07(0.50)**
Phish	94.18±2.05(0.00)	94.60±1.78(0.00)	93.48±1.64(0.00)	93.79±1.47(0.00)	**94.82±1.87(0.50)**	**94.80±1.83(0.23)**	94.78±1.85(0.03)	**94.80±1.88(0.23)**
Ionos	87.94±2.31(0.00)	88.34±2.25(0.04)	87.02±1.76(0.00)	87.56±2.05(0.00)	**88.80±2.18(0.50)**	88.38±2.03(0.03)	88.03±2.02(0.00)	**88.80±2.27(0.50)**
Wave	59.89±6.26(0.00)	74.95±5.43(0.01)	68.31±4.21(0.00)	67.58±4.24(0.00)	**75.25±5.56(0.50)**	74.91±5.48(0.00)	73.18±7.28(0.00)	**75.18±5.50(0.12)**
Conne	67.82±2.17(0.00)	70.59±2.47(0.00)	**74.41±3.74(0.50)**	72.88±3.17(0.00)	73.86±3.69(0.00)	72.76±3.39(0.00)	72.57±3.57(0.00)	74.20±3.52(0.02)
Nomao	91.91±1.71(0.00)	93.66±1.73(0.01)	92.87±1.59(0.00)	92.32±1.40(0.00)	93.50±1.76(0.00)	**93.75±1.75(0.37)**	92.84±2.61(0.00)	**93.76±1.78(0.50)**
Musk	75.35±4.61(0.00)	75.44±4.63(0.00)	**80.29±6.26(0.07)**	75.78±4.13(0.00)	77.83±4.30(0.00)	77.11±3.98(0.00)	76.88±4.17(0.00)	**80.84±7.78(0.50)**
Lung	82.58±7.76(0.00)	**87.01±8.31(0.21)**	70.74±4.93(0.00)	85.01±8.39(0.00)	**86.88±7.30(0.12)**	85.04±6.72(0.00)	76.06±9.80(0.00)	**87.45±7.38(0.50)**
UJIIn	94.43±3.43(0.00)	97.37±4.70(0.00)	97.98±4.27(0.00)	96.50±3.85(0.00)	96.92±4.75(0.00)	97.66±4.48(0.00)	96.87±4.78(0.00)	**98.12±4.27(0.50)**
Smart	81.70±3.65(0.00)	**88.72±7.01(0.29)**	**88.96±5.67(0.50)**	88.32±5.67(0.00)	87.78±7.01(0.01)	87.01±6.42(0.00)	70.43±6.14(0.00)	**88.38±6.55(0.07)**
Inter	94.39±0.31(0.00)	96.60±0.79(0.00)	95.28±0.54(0.00)	95.49±0.59(0.00)	96.66±0.90(0.01)	96.64±0.87(0.00)	96.39±0.80(0.00)	**96.72±0.89(0.50)**
Colon	79.55±5.25(0.00)	84.74±2.36(0.00)	**89.58±2.04(0.16)**	88.87±2.53(0.00)	87.61±2.96(0.00)	85.61±3.14(0.00)	81.00±2.42(0.00)	**90.03±2.48(0.50)**
SRBCT	66.16±5.50(0.00)	89.36±6.89(0.00)	73.80±9.36(0.00)	83.43±6.06(0.00)	85.16±6.39(0.00)	85.48±6.50(0.00)	**92.61±9.79(0.50)**	**92.46±9.57(0.43)**
DLBCL	83.16±4.31(0.00)	**96.80±5.91(0.50)**	85.75±5.37(0.00)	90.23±6.12(0.00)	94.32±6.89(0.00)	92.46±7.10(0.00)	94.32±6.65(0.00)	95.05±6.24(0.00)
TOX	59.79±3.69(0.00)	70.91±10.25(0.00)	73.85±6.60(0.00)	**78.74±7.23(0.50)**	70.59±6.17(0.00)	72.82±6.69(0.00)	60.40±3.58(0.00)	75.33±6.48(0.00)
Prost	84.94±2.38(0.00)	89.51±1.88(0.00)	87.59±2.56(0.00)	84.12±2.53(0.00)	89.61±3.00(0.00)	89.31±2.76(0.00)	87.61±2.21(0.00)	**91.31±2.65(0.50)**
Breast	65.76±6.18(0.00)	**88.14±6.97(0.44)**	63.45±7.32(0.00)	85.48±6.94(0.00)	86.48±6.71(0.00)	86.74±6.93(0.01)	83.69±6.00(0.00)	**88.21±5.96(0.50)**
Arcene	64.00±0.00(0.00)	70.20±6.05(0.00)	76.48±4.75(0.00)	76.48±4.75(0.00)	71.80±3.36(0.00)	72.32±3.79(0.00)	72.32±3.79(0.00)	**77.96±3.93(0.50)**
Cance	64.58±7.92(0.00)	64.69±14.98(0.00)	58.52±8.07(0.00)	75.21±10.33(0.00)	69.51±8.05(0.00)	**76.99±10.57(0.09)**	58.37±7.58(0.00)	**77.46±9.89(0.50)**
Leuk	89.08±5.14(0.00)	93.78±4.37(0.00)	88.47±5.45(0.00)	89.94±5.62(0.00)	92.97±4.58(0.00)	**95.22±5.03(0.17)**	88.64±4.55(0.00)	**95.78±5.53(0.50)**
GLI	85.58±2.53(0.00)	93.15±1.55(0.00)	79.13±12.04(0.00)	87.86±5.38(0.00)	92.45±1.49(0.00)	93.74±1.71(0.02)	91.58±1.45(0.00)	**94.28±1.78(0.50)**
GLA	59.33±4.66(0.00)	71.62±3.47(0.00)	61.02±7.04(0.00)	72.13±2.84(0.01)	71.73±3.45(0.00)	70.67±3.65(0.00)	67.26±3.19(0.00)	**73.14±3.62(0.50)**
AVG.	*78.29*	*85.11*	*80.94*	*84.62*	*85.40*	*85.62*	*82.21*	***87.43***
WIN	*0*	*4*	*4*	*1*	*5*	*5*	*1*	*19*

TABLE 3
Average SVM Classification Accuracy (Mean±Std.) with p-Value (in Percentage)

Data	Algorithm							
	MIFS	*mRMR*	*CIFE*	*ICAP*	*JMI*	*CMIM*	*Max-Rel*	*MRI*
Image	79.23±10.27(0.00)	84.49±10.59(0.00)	87.32±10.12(0.03)	**87.56±10.09(0.50)**	86.57±10.18(0.03)	**87.20±10.17(0.09)**	80.92±16.02(0.01)	**87.53±10.15(0.43)**
Phish	91.49±0.98(0.00)	**91.89±0.78(0.23)**	90.87±0.71(0.00)	90.80±0.68(0.00)	**91.89±0.83(0.28)**	**91.91±0.83(0.50)**	**91.91±0.81(0.40)**	**91.91±0.86(0.42)**
Ionos	**87.89±1.14(0.50)**	87.49±1.03(0.00)	83.53±1.18(0.00)	**87.74±1.80(0.30)**	86.77±1.84(0.00)	87.02±1.59(0.00)	86.30±1.65(0.00)	86.68±2.01(0.00)
Wave	74.18±6.15(0.00)	**84.52±5.79(0.50)**	80.36±4.78(0.00)	79.80±4.73(0.00)	**84.51±5.81(0.41)**	84.42±5.85(0.01)	83.23±7.30(0.00)	**84.48±5.81(0.16)**
Conne	64.78±0.00(0.00)	**65.83±0.00(0.50)**	64.78±0.00(0.00)	64.78±0.00(0.00)	64.78±0.00(0.00)	64.78±0.00(0.00)	64.78±0.00(0.00)	**65.83±0.00(0.50)**
Nomao	89.75±1.43(0.00)	**91.96±1.68(0.50)**	89.19±0.94(0.00)	88.93±0.91(0.00)	91.57±1.68(0.00)	**91.93±1.66(0.19)**	91.02±2.37(0.00)	91.76±1.68(0.00)
Musk	69.38±6.25(0.00)	69.45±6.20(0.00)	**70.53±7.42(0.50)**	68.00±3.64(0.00)	68.09±4.23(0.00)	**70.35±5.36(0.36)**	67.41±4.91(0.00)	**70.35±7.18(0.31)**
Lung	85.18±9.60(0.00)	**87.53±10.25(0.38)**	76.93±8.33(0.00)	83.64±9.20(0.00)	**87.67±10.58(0.50)**	**87.67±10.88(0.50)**	81.56±12.58(0.00)	**86.88±10.12(0.06)**
UJIIn	94.35±3.51(0.00)	97.23±4.79(0.00)	97.23±4.17(0.00)	96.43±3.92(0.00)	96.80±4.83(0.00)	97.54±4.59(0.00)	96.77±4.80(0.00)	**98.01±4.32(0.50)**
Smart	84.23±6.48(0.00)	85.06±8.49(0.00)	88.45±7.64(0.00)	**88.79±7.76(0.50)**	84.98±8.47(0.00)	87.31±8.48(0.00)	69.06±7.63(0.00)	86.34±8.22(0.00)
Inter	92.52±0.45(0.00)	**95.65±1.03(0.50)**	92.86±0.77(0.00)	93.10±0.73(0.00)	95.24±1.26(0.00)	95.53±1.38(0.04)	94.83±1.15(0.00)	95.43±1.36(0.00)
Colon	84.65±5.20(0.00)	84.97±2.45(0.00)	**88.23±1.73(0.50)**	86.45±3.51(0.00)	87.00±2.70(0.00)	86.71±2.56(0.00)	82.61±4.89(0.00)	86.84±3.93(0.01)
SRBCT	78.25±7.01(0.00)	88.61±12.97(0.00)	78.30±7.08(0.00)	86.43±12.85(0.00)	85.61±13.45(0.00)	84.11±12.89(0.00)	**91.55±11.90(0.12)**	**92.82±8.56(0.50)**
DLBCL	87.59±8.82(0.00)	92.93±9.02(0.00)	79.32±11.90(0.00)	88.05±11.84(0.00)	92.50±10.68(0.00)	90.25±10.31(0.00)	90.86±10.01(0.00)	**94.82±9.36(0.50)**
TOX	67.80±4.25(0.00)	66.57±7.74(0.00)	68.40±4.15(0.00)	77.60±6.83(0.04)	73.68±6.24(0.00)	**78.32±7.47(0.29)**	66.21±4.78(0.00)	**78.55±8.10(0.50)**
Prost	90.39±1.99(0.00)	92.47±1.96(0.00)	90.29±1.94(0.00)	92.04±2.37(0.00)	**93.39±2.12(0.19)**	93.22±2.14(0.04)	92.82±1.71(0.00)	**93.47±2.08(0.50)**
Breast	74.07±5.58(0.00)	94.02±9.58(0.00)	68.41±3.62(0.00)	87.91±9.69(0.00)	89.50±7.87(0.00)	88.91±10.71(0.00)	88.38±6.65(0.00)	**96.00±8.01(0.50)**
Arcene	56.00±0.00(0.00)	68.98±5.07(0.00)	**71.44±5.17(0.50)**	**71.44±5.17(0.50)**	65.30±4.75(0.00)	64.92±4.53(0.00)	64.98±4.59(0.00)	68.60±4.95(0.00)
Cance	76.95±9.42(0.00)	65.83±15.76(0.00)	55.44±4.13(0.00)	77.71±10.81(0.00)	75.33±11.08(0.00)	78.40±10.18(0.04)	61.43±8.90(0.00)	**79.13±10.21(0.50)**
Leuk	90.36±3.91(0.00)	95.86±5.22(0.01)	72.47±4.38(0.00)	93.92±5.12(0.00)	95.42±5.40(0.00)	96.09±6.21(0.01)	92.61±6.81(0.00)	**96.86±6.59(0.50)**
GLI	93.58±1.78(0.00)	94.26±3.37(0.00)	88.49±5.03(0.00)	**96.26±3.33(0.46)**	95.06±2.18(0.00)	**96.31±2.13(0.50)**	90.87±1.53(0.00)	95.95±1.78(0.03)
GLA	75.57±4.16(0.00)	78.88±5.56(0.03)	68.30±2.52(0.00)	**79.88±6.38(0.50)**	77.32±4.69(0.00)	77.97±4.75(0.00)	70.72±4.78(0.00)	78.08±4.51(0.01)
AVG.	*81.28*	*84.75*	*79.60*	*84.88*	*84.95*	*85.49*	*81.86*	***86.65***
WIN	*1*	*6*	*3*	*6*	*4*	*7*	*2*	*14*

on some single data sets. For instance, on the data sets of Image, Phish, Conne, Musk, Lung, UJIIn, Inter, and Arcene, the compared criteria perform better on the 1-NN classifier than on the SVM classifier. Yet the compared criteria achieve better performance on the SVM classifier on the data sets of Wave, Prost, Breast and GLA. Generally speaking, it follows from Tables 2 and 3 that *MRI* is comparable or superior to the other mutual information-based criteria. *mRMR*, *JMI*, and *CMIM* also perform well, although not better than *MRI*.

To further illustrate the classification performance, the variations of average classification accuracies of the 1-NN classifier and SVM classifier on nine benchmark data sets in Table 1 are depicted in Fig. 9. The number of selected features increases from 5 to 50 in the interval of 5 (on the data sets of Waveform and Connect, it reaches up to 40). Four baseline evaluation criteria, *mRMR*, *CIFE*, *JMI*, and *Max_Rel*, are compared with *MRI*, which are the representative redundancy reduction criteria, new information maximization criterion, and top-k criterion, respectively.

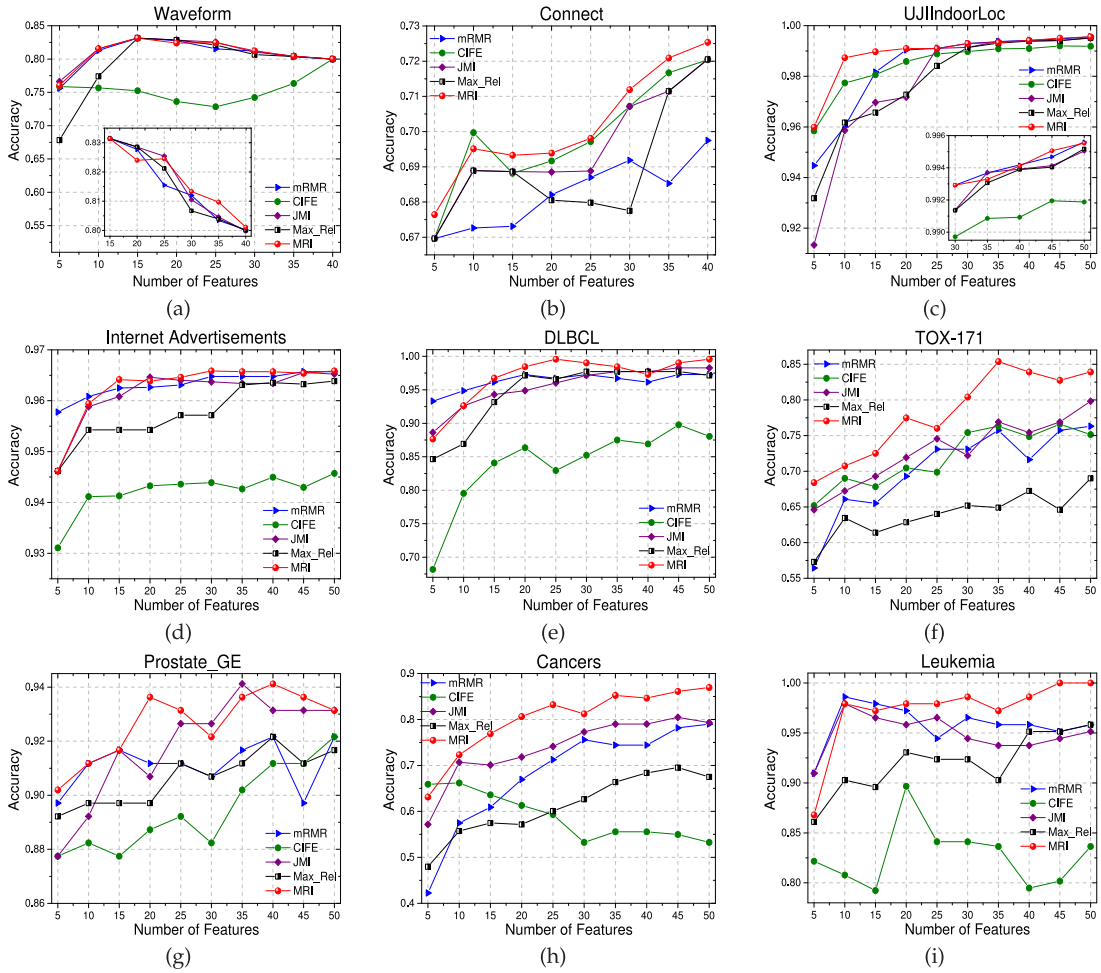


Fig. 9. Average classification accuracy of the 1-NN classifier and SVM classifier.

As shown in Fig. 9, classification accuracy exhibits different variations across different data sets. On some data sets, such as Waveform, UJIIndoorLoc, Internet, and DLBCL, some criteria achieve their best performance only with a few features. For this kind of data sets, convergence state would be satisfied within a small range of features if adopting a classical wrapper model to select features[6]. In other words, the classification performance for these data sets is not enhanced by monotonously expanding the number of selected features, sometimes may be weakened, such as the case of Waveform in Fig. 9. Conversely, classification accuracy exhibits a rising trend on other data sets when more features are selected. Therefore, how to determine the optimal size of feature subset is still an open question for filter selection approaches with individual feature evaluation strategy. Generally speaking, *MRI* outperforms or is comparable with the other criteria in its ability of selecting discriminative features as demonstrated in Fig. 9.

Other metrics, i.e., Balanced Error Rate (BER), Area Under ROC Curve (AUC), Kuncheva's Stability Index (Stability) [51], and Inconsistency Rate [47], are also employed to evaluate the performance of feature subsets, as shown in Fig. 10. The size of feature subsets increases from 5 to 50 in the interval of 5, and the average BER, AUC, stability, and inconsistency rate across all of the benchmark data sets are represented by boxes in Fig. 10. The boxes indicate median values and lower and upper quartiles. Additionally, minimal, mean and maximal values are represented by red crossbars in Fig. 10. BER concerns the unbalanced distributions of classes and counts the average classification error for each class. AUC exploits a trapezoidal method to estimate the area under ROC curve, and is widely used to assess the discriminative performance. Stability measures the consistency of pair-wise feature subsets to demonstrate how stable an approach performs in selection process. In contrast, inconsistency rate measures the inconsistency of pair-wise patterns in an individual feature subset. That is to say, the four metrics evaluate the performance of the selected feature subsets from different aspects.

Overall, as illustrated in Fig. 10, *MRI* is better or comparable with the other criteria in most cases under the metrics of BER, AUC, stability, and inconsistency rate. Note that *MRI* is a little inferior to *Max-Rel* under the metric of stability. This mainly attributes to the characteristic of the stability index, i.e., it does not consider feature redundancy. The feature subset consisting of many relevant but redundant features is highly scored by this metric. Thus, *Max-Rel* performs best among all of the compared mutual information-based criteria, and is also better than *MRI* that alleviates feature redundancy in the selected subset. Generally, *JMI* and *CMIM* also show comparably better than the other criteria except *MRI*. That is, these two criteria also have excellent selection abilities. This is in some sense coincident with the observation reported in [26].

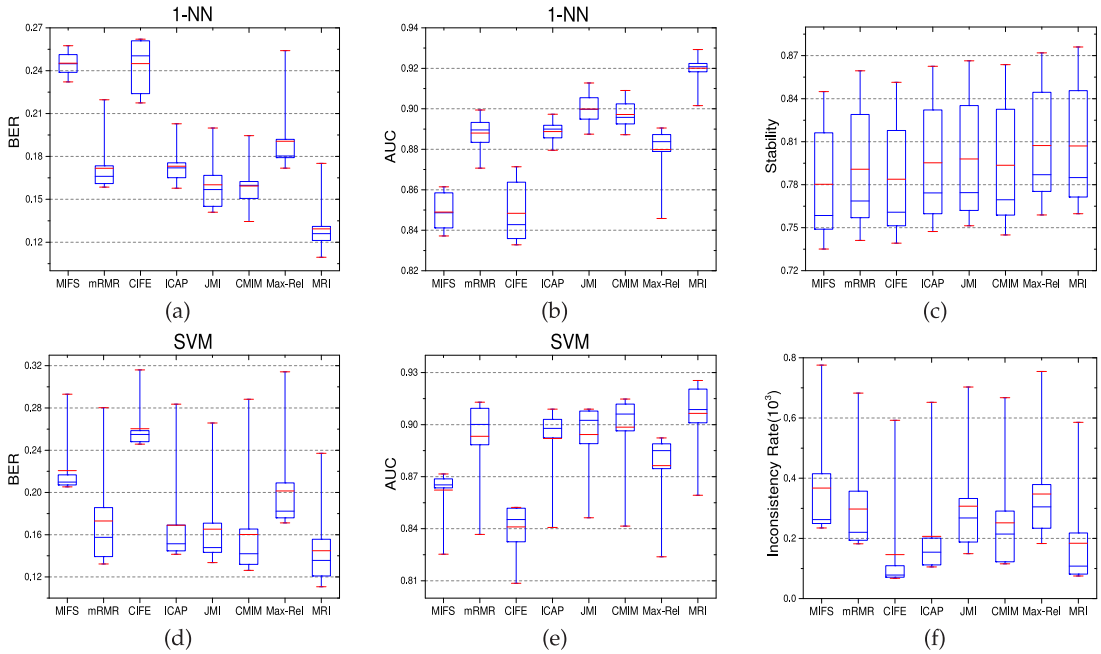


Fig. 10. Average BER (the lower of BER, the better), AUC (the higher of AUC, the better), Stability (the higher of the stability, the better), and inconsistency rate (the lower of the inconsistency rate, the better) across all of the benchmark data sets.

5.4 Exp. 4: Analyzing Parameter Sensitivity

In the above experiments, the parameter β of *MIFS* is fixed to 1, according to the suggestion of the author [11]. β plays a role of balancing the importance of the classification ability of an individual feature and its redundancy with the selected subset. The optimal β is complicated to be determined across different data sets, although *MIFS* is well-known for its capability of reducing feature redundancy and used as a baseline criterion in many works [13], [15], [16], [26], [40], [43]. In this section, the issue of parameter sensitivity is discussed between *MIFS* and *MRI*.

Four benchmark data sets in Table 1, i.e., Image, Musk, UJIIndoorLoc, and Cancers, are tested. On Image, *MIFS* and *MRI* sequentially select top-34 features for a comparison. The number of selected features, i.e., k, sequentially increases from 1 to 34. On Musk, UJIIndoorLoc, and Cancers, k reaches up to 50. The balance parameter β of *MIFS* is tuned from 0 to 1 in the interval of 0.1. The average SVM classification accuracy of *MIFS* on each benchmark data set is demonstrated in Fig. 11, where the accuracy of *MRI* is also depicted.

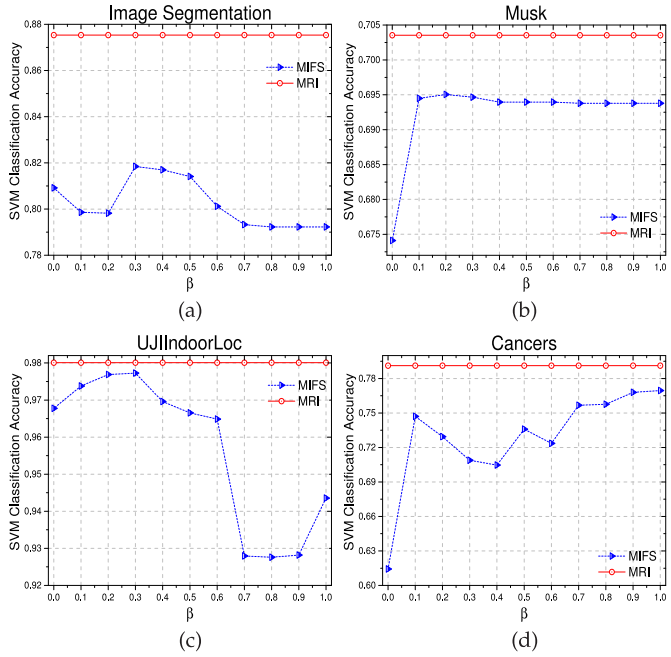


Fig. 11. Average SVM classification accuracy of *MIFS* and *MRI* when β increases.

Note that *Max-Rel* criterion is actually performed when β is equal to 0. As suggested in Fig. 11, the performance of *MIFS* is closely related to β. On Image and UJIIndoorLoc, *MIFS* achieves its best performance when $\beta = 0.3$. Yet on Cancers, it yields the best result at the value of 1. An interesting case appears on Musk, that is, the classification performance of *MIFS* shows a slight variation through the process of tuning β. In contrast, *MRI* balances its redundancy term and new classification information term with equal weight, rather than tuning their importance to search an optimal solution.

5.5 Exp. 5: Comparing with Non-Mutual-Information-Based Feature Selection Approaches

In this section, *MRI* is compared with several popular non-mutual-information-based feature selection approaches, i.e., *Laplacian Score* [30], *Inf-FS* [31], *ReliefF* [32], *SPEC* [33], and *SPFS* [34]. The affinity matrix in *Laplacian Score*, *SPEC* and *SPFS* is calculated as $\mathbf{K}_{ij} = \begin{cases} a_{ij}, & \mathbf{y}_i = \mathbf{y}_j \\ 0, & otherwise \end{cases}$, where a_{ij} is the similarity between the instances $\mathbf{I}_i$ and $\mathbf{I}_j$ and determined by the *RBF* kernel function: $a_{ij} = \exp(-\frac{\|\mathbf{I}_i - \mathbf{I}_j\|^2}{2\delta^2})$, $\delta^2 = mean(\|\mathbf{I}_i - \mathbf{I}_j\|^2)$. *SPEC* adopts the $\hat{\varphi}_1(\cdot)$ ranking function according to the literature [33]. *SPFS* is implemented through SFS. The number of the nearest instances in *ReliefF* is set to 10, and all of the instances are sampled in evaluation. For *Inf-FS*, its loading coefficient is fixed to 0.5.

The size of the selected feature subset, i.e., k, increases from 10 to 50 in the interval of 10, and its performance is evaluated under the metrics of accuracy (ACC), BER, and AUC,

TABLE 4
Average ACC, BER, and AUC (Mean±Std.) with p-Value for the 1-NN Classifier and SVM Classifier Across All of the Benchmark Data Sets (in Percentage): k Represents the Number of Selected Features

#k	Metric	Algorithm					
		Laplacian	*Inf-FS*	*ReliefF*	*SPEC*	*SPFS*	*MRI*
10	ACC	73.74±19.69(0.00)	68.44±17.63(0.00)	82.10±13.48(0.01)	76.87±16.58(0.00)	83.25±9.86(0.04)	**86.66±9.54(0.50)**
	BER	30.21±21.88(0.00)	38.14±20.75(0.00)	22.25±16.97(0.01)	27.72±18.68(0.00)	**20.37±12.45(0.05)**	**15.29±11.50(0.50)**
	AUC	80.82±13.65(0.00)	74.34±13.97(0.00)	86.83±9.20(0.02)	82.41±12.18(0.00)	87.17±7.40(0.02)	**90.69±8.12(0.50)**
20	ACC	77.98±19.34(0.00)	72.70±14.84(0.00)	85.14±11.86(0.01)	82.32±12.37(0.00)	81.83±10.68(0.00)	**89.20±9.05(0.50)**
	BER	25.56±20.35(0.00)	32.91±16.95(0.00)	18.64±14.13(0.01)	21.00±12.65(0.00)	21.26±12.09(0.00)	**13.52±10.09(0.50)**
	AUC	82.84±13.60(0.00)	77.68±13.02(0.00)	87.68±9.39(0.01)	86.38±8.36(0.00)	85.65±7.26(0.00)	**91.27±8.14(0.50)**
30	ACC	79.36±18.98(0.00)	76.72±13.08(0.00)	86.25±10.39(0.01)	83.87±11.18(0.00)	80.77±10.93(0.00)	**88.83±8.04(0.50)**
	BER	23.96±19.86(0.00)	28.00±14.28(0.00)	16.99±11.64(0.01)	19.45±12.42(0.00)	22.23±12.70(0.00)	**13.29±8.62(0.50)**
	AUC	84.12±13.67(0.00)	80.53±12.17(0.00)	88.92±8.28(0.01)	86.93±8.40(0.00)	84.62±8.04(0.00)	**91.34±7.08(0.50)**
40	ACC	79.66±19.10(0.00)	79.91±13.01(0.00)	87.08±9.22(0.00)	83.03±11.48(0.00)	78.71±11.24(0.00)	**89.64±7.95(0.50)**
	BER	23.61±19.91(0.00)	23.29±12.89(0.00)	16.06±10.39(0.01)	20.32±12.36(0.00)	24.04±12.71(0.00)	**12.42±8.35(0.50)**
	AUC	84.28±13.96(0.00)	83.47±12.12(0.00)	90.07±7.36(0.03)	87.02±8.30(0.00)	83.56±8.71(0.00)	**91.92±6.50(0.50)**
50	ACC	80.01±19.25(0.00)	81.30±11.92(0.00)	87.82±8.80(0.00)	85.09±10.24(0.00)	78.57±12.02(0.00)	**91.13±7.66(0.50)**
	BER	23.46±20.57(0.00)	21.94±11.74(0.00)	15.13±9.83(0.01)	18.11±11.54(0.00)	24.09±13.35(0.00)	**11.23±7.93(0.50)**
	AUC	84.87±14.15(0.00)	84.18±12.11(0.00)	90.27±7.33(0.02)	88.51±7.76(0.00)	83.32±9.37(0.00)	**92.57±6.56(0.50)**
AVG.	ACC	*78.15*	*75.81*	*85.68*	*82.24*	*80.63*	***89.09***
	BER	*25.36*	*28.86*	*17.81*	*21.32*	*22.40*	***13.15***
	AUC	*83.37*	*80.04*	*88.75*	*86.25*	*84.86*	***91.56***

as recorded in Table 4. It shows that *MRI* yields excellent results under these metrics. *ReliefF* also shows relatively nice performance, although not better than *MRI*. It assesses discriminative abilities of features by counting the difference between sampled instance and its nearest neighbors respectively in the same and different classes. Yet in essence, *ReliefF* is a dependency-based approach, which disregards feature redundancy in its evaluation criterion. Amount of redundant classification information in the selected feature subset results in its inferior performance to *MRI*. Note that the loading coefficient of *Inf-FS* is fixed through the experiments. The performance of *Inf-FS* would commonly be enhanced if this parameter is deliberately tuned, while the time consumption would increase accordingly.

6 Conclusion

A new mutual information term, namely, independent classification information, is defined in this paper. It encompasses both the independent information that a candidate feature provides and the independent information that the selected features preserve. Independent classification information is proved as a loose upper bound of the total classification information of feature subset. Thus, the maximization of independent classification information helps enhance the global discriminative performance. Then, a new feature evaluation criterion, i.e., *MRI*, is proposed on the basis of independent classification information. Besides pursuing the maximization of feature relevance with classes, *MRI* maximizes independent classification information. By analysis and comparison with some popular evaluation criteria, *MRI* is illustrated to properly regulate the effects of feature relevance and feature redundancy, neither of which is exaggerated or depreciated in estimating the contribution of feature to classification. Comprehensive experiments on various data sets testify the effectiveness of *MRI* in selecting highly predictive and lowly redundant features.

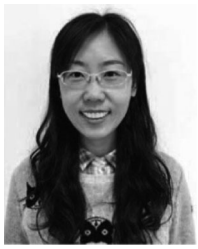
Acknowledgments

The authors are very grateful to the anonymous reviewers and editor for their helpful and constructive comments and suggestions. This work is partially supported by the National Natural Science Foundation of China under Grant No. 61070089 and No. 11431006, the Natural Science Foundation of Tianjin City under Grant No. 14JCYBJC15700 and No. 15JCYBJC46600, the Ministry of Education of Humanities and Social Science Project under Grant No. ZX20160077, the Research Fund for International Young Scientists under Grant No. 61650110510, and the Joint Fund of NSFC-Basic Research Institute of General Technology under Grant No. U1636116.

References

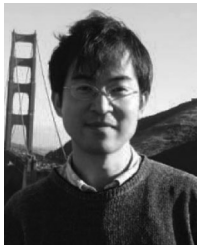
[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 1991.
[2] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clustering comparison: Variants, properties, normalization and correction for chance," *J. Mach. Learn. Res.*, vol. 11, pp. 2837–2854, 2010.
[3] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.
[4] N. X. Vinh, J. Chan, S. Romano, and J. Bailey, "Effective global approaches for mutual information based feature selection," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 512–521.
[5] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artif. Intell.*, vol. 97, no. 1, pp. 245–271, 1997.
[6] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, no. 1, pp. 273–324, 1997.
[7] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature Extraction: Foundations and Applications*. Berlin, Germany: Springer-Verlag, 2006, ch. 6.
[8] L. Breiman, "Probability," in *Classics in Applied Mathematics*, vol. 7. Philadelphia, PA, USA: SIAM, 1992.
[9] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *J. Mach. Learn. Res.*, vol. 5, pp. 1205–1224, 2004.
[10] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. Boston, MA, USA: Kluwer, 1998.
[11] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Trans. Neural Netw.*, vol. 5, no. 4, pp. 537–550, Jul. 1994.
[12] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.
[13] P. A. Estévez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," *IEEE Trans. Neural Netw.*, vol. 20, no. 2, pp. 189–201, Feb. 2009.

[14] D. Lin and X. Tang, "Conditional infomax learning: An integrated framework for feature extraction and fusion," in *Proc. 9th Eur. Conf. Comput. Vis.*, 2006, pp. 68–82.
[15] N. Kwak and C. H. Choi, "Input feature selection for classification problems," *IEEE Trans. Neural Netw.*, vol. 13, no. 1, pp. 143–159, Jan. 2002.
[16] H. Cheng, Z. Qin, C. Feng, Y. Wang, and F. Li, "Conditional mutual information-based feature selection analyzing for synergy and redundancy," *Electron. Telecommun. Res. Inst. J.*, vol. 33, no. 2, pp. 210–218, 2011.
[17] A. Jakulin, "Machine learning based on attribute interactions," Ph.D. dissertation, Faculty Comput. Inf. Sci., Ljubljana Univ., Ljubljana, Slovenia, 2005.
[18] G. Bontempi and P. E. Meyer, "Causal filter selection in microarray data," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 95–102.
[19] A. E. Akadi, A. E. Ouardighi, and D. Aboutajdine, "A powerful feature selection approach based on mutual information," *Int. J. Comput. Sci. Netw. Secur.*, vol. 8, no. 4, pp. 116–121, 2008.
[20] R. W. Yeung, "A new outlook on Shannon's information measures," *IEEE Trans. Inf. Theory*, vol. 37, no. 3, pp. 466–474, May 1991.
[21] J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," *Neural Comput. Appl.*, vol. 24, no. 1, pp. 175–186, 2014.
[22] H. Yang and J. Moody, "Data visualization and feature selection: New algorithms for nongaussian data," *Advances Neural Inf. Process. Syst.*, vol. 12, pp. 687–693, 1999.
[23] M. Vidal-Naquet and S. Ullman, "Object recognition with informative features and linear classification," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, 2003, pp. 281–288.
[24] P. E. Meyer and G. Bontempi, "On the use of variable complementarity for feature selection in cancer classification," in *Applications of Evolutionary Computing*. Berlin, Germany: Springer, 2006, pp. 91–102.
[25] F. Fleuret, "Fast binary feature selection with conditional mutual information," *J. Mach. Learn. Res.*, vol. 5, pp. 1531–1555, 2004.
[26] G. Brown, A. Pocock, M. Zhao, and M. Luján, "Conditional likelihood maximisation: A unifying framework for information theoretic feature selection," *J. Mach. Learn. Res.*, vol. 13, pp. 27–66, 2012.
[27] K. Bache and M. Lichman, "UCI machine learning repository," Univ. of California, School Inf. Comput. Sci., Irvine, 2013. [Online]. Available: http://archive.ics.uci.edu/ml
[28] H. Peng, "Mutual information computation," 2007. [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/14888-mutual-information-computation
[29] J. Tang, S. Alelyani, and H. Liu, "Feature felection for classification: A review," in *Data Classification: Algorithms and Applications*. Chapman, CA, USA: CRC Press, 2014.
[30] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Proc. Advances Neural Inf. Process. Syst.*, 2005, pp. 507–514.
[31] G. Roffo, S. Melzi, and M. Cristani, "Infinite feature selection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4202–4210.
[32] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Mach. Learn.*, vol. 53, no. 1/2, pp. 23–69, 2003.
[33] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 1151–1157.
[34] Z. Zhao, L. Wang, H. Liu, and J. Ye, "On similarity preserving feature selection," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 3, pp. 619–632, Mar. 2013.
[35] Q. Song, J. Ni, and G. Wang, "A fast clustering-based feature subset selection algorithm for high dimensional data," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 1, pp. 1–14, Jan. 2013.
[36] N. Hoque, D. K. Bhattacharyya, and J. K. Kalita, "MIFS-ND: A mutual information-based feature selection method," *Expert Syst. Appl.*, vol. 41, pp. 6371–6385, 2014.
[37] N. X. Vinh and J. Bailey, "Comments on supervised feature selection by clustering using conditional mutual information-based distances," *Pattern Recognit.*, vol. 46, no. 4, pp. 1220–1225, 2013.
[38] A. Jakulin and I. Bratko, "Quantifying and visualizing attribute interactions," 2003. [Online]. Available: arXiv: cs/0308002
[39] P. E. Meyer, C. Schretter, and G. Bontempi, "Information-theoretic feature selection in microarray data using variable complementarity," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 3, pp. 261–274, Jun. 2008.
[40] J. M. Sotoca and F. Pla, "Supervised feature selection by clustering using conditional mutual information-based distances," *Pattern Recognit.*, vol. 43, no. 6, pp. 2068–2081, 2010.
[41] D. Koller and M. Sahami, "Toward optimal feature selection," in *Proc. 13th Int. Conf. Mach. Learn.*, 1996, pp. 284–292.
[42] R. M. Fano and D. Hawkins, "Transmission of information: A statistical theory of communications," *Amer. J. Phys.*, vol. 29, no. 11, pp. 793–794, 1961.
[43] N. Kwak and C. H. Choi, "Input feature selection by mutual information based on Parzen window," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 12, pp. 1667–1671, Dec. 2002.
[44] K. Torkkola, "Feature extraction by non-parametric mutual information maximization," *J. Mach. Learn. Res.*, vol. 3, pp. 1415–1438, 2003.
[45] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Exploration Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.
[46] P. Mitra, C. A. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 301–312, Mar. 2002.
[47] M. Dash and H. Liu, "Consistency-based search in feature selection," *Artif. Intell.*, vol. 151, no. 1, pp. 155–176, 2003.
[48] Q. Gu, Z. Li, and J. Han, "Generalized Fisher score for feature selection," in *Proc. 27th Conf. Uncertainty Artif. Intell.*, 2011, pp. 266–273.
[49] J. M. Wei, S. Q. Wang, and X. J. Yuan, "Ensemble rough hypercuboid approach for classifying cancers," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 3, pp. 381–391, Mar. 2010.
[50] J. Li, et al., "Feature selection: A data perspective," *arXiv:1601.07996*, 2016.
[51] L. I. Kuncheva, "A stability index for feature selection," in *Proc. 25th IASTED Int. Multi-Conf. Artif. Intell. Appl.*, 2007, pp. 390–395.



Jun Wang received the MS degree in computer science from Shandong Normal University, China, in 2007. Now, she is working toward the PhD degree in the College of Computer and Control Engineering, Nankai University, China. From 2007 to 2013, she worked as a lecturer in the College of Mathematics and Statistics Science, Ludong University, China. Her research interests include data mining and machine learning.

Jin-Mao Wei received the PhD degree from the East China University of Science and Technology, China, in 2001. From 2003 to 2006, he worked as a postdoctoral fellow in the College of Computer Science and Technology, Jilin University, China. Before 2007, he was with North-east Normal University. He is now a professor in the College of Computer and Control Engineering, Nankai University, China. His current research works include machine learning, data mining, Web mining, and bioinformatics.



Zhenglu Yang received the PhD degree from the University of Tokyo, Japan, in 2008. From 2008 to 2014, he worked as a faculty in the Institute of Industrial Science, University of Tokyo. He is now a professor in the College of Computer and Control Engineering, Nankai University, China. His research interests include artificial intelligence, database systems, and big data mining.

Shu-Qin Wang received the PhD degree from Jilin University, China, in 2009. She is now an associate professor in the College of Computer and Information Engineering, Tianjin Normal University, China. Her current research interests include data mining and bioinformatics.