Evaluating and Selecting Features via Information Theoretic Lower Bounds of Feature Inner Correlations for High-Dimensional Data

Yishi Zhang[a], Ruilin Zhu[b], Zhijun Chen[c,*], Jie Gao[d], De Xia[a,*]

[a]*School of Management, Wuhan University of Technology, Wuhan 430070, China*
[b]*Management School, Lancaster University, Lancaster, LA1 4YX, United Kingdom*
[c]*Intelligent Transportation Systems Research Center, Wuhan University of Technology, Wuhan 430063, China*
[d]*School of Business Administration, Southwestern University of Finance and Economics, Chengdu 611130, China*

Abstract

Feature selection is an important preprocessing and interpretable method in the fields where big data plays an essential role. In this paper, we first reformulate and analyze some representative information theoretic feature selection methods from the perspective of approximations of feature inner correlations, and indicate that many of these methods cannot guarantee any theoretical bounds of feature inner correlations. We thus introduce two lower bounds that have very simple forms for feature redundancy and complementarity, and verify that they are closer to the optima than the existing lower bounds applied by some state-of-the-art information theoretic methods. A simple and effective feature selection method based on the proposed lower bounds is then proposed and empirically verified with a wide scope of real-world datasets. The experimental results show that the proposed method achieves promising improvement on feature selection, indicating the effectiveness of the feature criterion consisting of the proposed lower bounds of redundancy and complementarity.

Keywords: Data mining; feature selection; redundancy; complementarity; lower bounds

*Corresponding authors. Zhijun Chen serves as the first corresponding author of this paper.
Email addresses: easezh@126.com (Yishi Zhang), ruilin.zhu@lancaster.ac.uk (Ruilin Zhu), chenzj556@whut.edu.cn (Zhijun Chen), karina-gj@foxmail.com (Jie Gao), xiadetiger@126.com (De Xia)

1. Introduction

Big data have become common with the ongoing developments in information technology in recent years. Analyzing big data is a key task in data mining and machine learning techniques in the fields of business intelligence (e.g., Maldonado et al., 2015, 2017; López and Maldonado, 2019), causal discovery and inference (e.g., Ling et al., 2020; Yu et al., 2020), information retrieval (e.g., Yu et al., 2003), image processing (e.g., Chang et al., 2014), and bioinformatics (e.g., Antonov et al., 2004; Chen and Chen, 2009), among others. However, the substantial noise in big data has formed barriers for the learning algorithms to effectively and efficiently discriminate different classes. Feature selection, one of the typical dimensionality reduction techniques that selects representative features with good enough discriminative power for data representation and modeling, has been widely considered to be an essential preprocessing approach because it can effectively reduce the data acquisition and storage demands, increase learning speeds, and improve generalization performance (Boln-Canedo and Alonso-Betanzos, 2019; Das et al., 2020). It is also deemed as useful in finding the direct causes and effects for data-driven causal discovery (Ling et al., 2020) and in enhancing model interpretation (Chen et al., 2018).

Generally speaking, feature selection methods can be divided into three types: wrappers, embedded methods, and filters. Wrappers utilize the results of a specific learning algorithm to select features. The performance of the wrapper method is affected by the learning algorithm it used. Also, wrapper method tends to overfit on small training sets (Brown et al., 2012). For embedded methods, feature selection is integrated into classification process (Ghaddar and Naoum-Sawaya, 2018), therefore they are also overly specific to the classifier used. Filters apply the classifier-irrelevant metrics like Fisher score (Furey et al., 2000), χ^2-test (Qu et al., 2005), mutual information (MI) (Bennasar et al., 2013, 2015) to estimate the discriminative power of features. Thus the computational cost of them are often lower than wrappers and embedded methods. In addition to its generality, filters is an appropriate choice when dealing with high dimensional data. Typical filters include information theoretic feature selection methods, e.g., the methods using mutual information maximization (MIM) criterion (Lewis, 1992) and using minimum redundancy and maximum relevance(mRMR) criterion (Peng et al., 2005), have been widely used in many fields because of their excellent performance.

Despite the aforementioned information theoretic feature selection methods, there is another stream of research using optimization techniques to handle feature selection, wherein typical examples are those with maximum class separation distance criterion (Antonov et al., 2004), generalized Benders decomposition (Aytug, 2015), greedy randomized adaptive search procedure (Bertolazzi et al., 2016), convex semi-infinite programming (Won et al., 2020), and sparse representation (Bertsimas et al., 2020). Optimization technique-based methods features the novelty of formulating feature selection as an analytically solvable optimization task. However, solving some of the goal functions, e.g., those consisting of the ℓ_p-norm ($0 \leq p < 1$) regular-

izer and those consisting of the discrete optimization styles, are proved to be **NP**-hard (Chen et al., 2010). Although there exist several heuristic proxies, e.g., ℓ_1-norm (namely, LASSO; Tibshirani, 1996) or other convex approximations, matrix computation still incurs excessive execution time and space cost, causing hindrance to the implementation of such methods on large-scale datasets. In addition, optimization-based feature selection does not explicitly handle feature correlations, and thus makes it less interpretable and feasible to precisely determine whether or not sufficient efforts have been undertaken to handle feature correlations (Zhang et al., 2019). In contrast, information theoretic methods explicitly decompose the objective of feature selection into multiple sub-objectives (e.g., maximizing class-relevance and minimizing redundancy), and then apply flexible searching strategies for those sub-objectives to finally obtain "satisfactory" solutions. It is thus deemed as the method that can obtain efficient and interpretable results (i.e., facilitating a better understanding for the learning model or data) (Chen et al., 2018). Furthermore, MI features the following advantages: (a) it is related to the lower bound of the Bayes prediction error (Fano, 1961), and (b) it can be estimated highly efficiently (Zhang et al., 2019). As such, in this study, we focus on information theoretic feature selection.

The existing information theoretic feature selection methods focus not only on class-relevance (i.e., relevance between the feature and the class) but also on feature inner correlations (hereafter *inner correlations* refer to the correlations among the features, and *outer correlation* refers to the class-relevance) to explicitly eliminate redundancy (Bennasar et al., 2015; Liu et al., 2015; Gao et al., 2020). For instance, methods based on the mRMR criterion (Peng et al., 2005; Ding and Peng, 2003) and fast correlation-based feature selection algorithm (Yu and Liu, 2004; Song et al., 2013) are the most representative methods for redundancy elimination. Although redundancy is extensively considered by the above methods, another feature inner correlation, called complementarity which can improve the classification accuracy, does not receive much explicit attention in the literature. The conditional mutual information maximization (CMIM) criterion (Fleuret, 2004) actually considers the complementarity implicitly using the conditional mutual information (CMI). Other typical criteria, such as those proposed by Zhang et al. (2014) and Wang et al. (2017), are also built based on CMI. Complementarity can also be implicitly identified using joint mutual information (JMI) (Yang and Moody, 1999; Guo and Nixon, 2009; Bennasar et al., 2015), which is actually the MI between the feature group (the existing works focus primarily on pairs of features) and the class. Several studies, e.g., Brown et al. (2012) and Wang et al. (2017), consciously expand JMI into three dimensions, namely, class-relevance, redundancy, and conditional redundancy, and explicitly handle redundancy and complementarity using the latter two terms.

Although these feature selection methods aim to select salient features according to no more than the three aforementioned dimensions of feature correlations, most of them apply hand-designed heuristic criteria with pairwise approximations and cannot guarantee theoretical bounds for the higher-order inner correlations of features. Departing from the linear span of possible information theoretic feature selection criteria

proposed by Brown et al. (2012), this paper focuses on the nonlinear one, and attempts to select features via the lower bounds of redundancy and complementarity within the nonlinear span. The main contributions of this paper are as below:

(i) The representative information theoretic feature selection methods are reformulated and analyzed in the context of the three dimensions of feature correlations, i.e., the class-relevance, redundancy, and the complementarity.

(ii) The weighted pairwise redundancy and complementarity are deemed as the approximations of their higher-order forms.

(iii) The linear span of the information theoretic feature selection criteria regarding the three dimensions is extended to the nonlinear one.

(iv) A criterion with loose lower bounds from the nonlinear span for redundancy and complementarity is proposed and empirically verified to be effective in feature selection.

The remainder of this paper is organized as follows. Section 2 presents the information theoretic metrics and the theoretical foundation of applying MI for feature selection. Section 3 reformulates and analyzes the existing representative methods as well as the potential ones within a unified framework, and discusses the extension of the linear span of the information theoretic feature selection criteria regarding the class-relevance, redundancy, and the complementarity. Section 4 proposes two loose lower bounds with very simple forms from the nonlinear span for redundancy and complementarity, respectively, and verifies that they are closer to the optima compared with the existing bounds utilized by the state-of-the-art information theoretic methods. A simple and effective feature selection method based on the proposed bounds is then proposed. Section 5 empirically verifies the effectiveness of the proposed method via a wide scope of real-world datasets. Section 6 thereafter concludes this study and proposes possible further work.

2. Preliminaries

Some essential metrics in information theory will be briefly introduced in this section. Three dimensions of feature correlations, i.e., class-relevance, redundancy, and complementarity, will then be introduced and analyzed. Finally, the representative information theoretic feature selection methods will be reformulated and analyzed from the perspective of the bounds of feature inner correlations.

2.1. Information theoretic metrics

MI and CMI are the most frequently used metrics in feature selection methods (Brown et al., 2012). The MI between two random variable sets $\mathbf{X} = \{X_1, ..., X_k\}$ and $\mathbf{Y} = \{Y_1, ..., Y_l\}$ can be expressed as follows (Cover and Thomas, 1991):

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

where $\mathbf{x} \in \mathbf{X}$ and $\mathbf{y} \in \mathbf{Y}$ are the possible value assignments of $\mathbf{X}$ and $\mathbf{Y}$, respectively. MI can be considered as the amount of information shared by two variable sets. In the field of feature selection, MI is one of the most widely used metrics for measuring the correlation intensity of two features. MI is a symmetrical metric, i.e., $I(\mathbf{X};\mathbf{Y}) = I(\mathbf{Y};\mathbf{X})$. $I(\mathbf{X};\mathbf{Y}) = 0$ indicates independence between $\mathbf{X}$ and $\mathbf{Y}$, while $I(\mathbf{X};\mathbf{Y}) >> 0$ indicates strong dependence.

CMI, an extension of MI for measuring the conditional dependence between two variable sets given a third set, is defined as (Cover and Thomas, 1991)

$$I(\mathbf{X};\mathbf{Y}|\mathbf{Z}) = \sum_{\mathbf{z}\in\mathbf{Z}} p(\mathbf{z}) \sum_{\mathbf{x}\in\mathbf{X}} \sum_{\mathbf{y}\in\mathbf{Y}} p(\mathbf{xy}|\mathbf{z}) \log \frac{p(\mathbf{xy}|\mathbf{z})}{p(\mathbf{x}|\mathbf{z})p(\mathbf{y}|\mathbf{z})}.$$

$I(\mathbf{X};\mathbf{Y}|\mathbf{Z})$ can be interpreted as the information shared between $\mathbf{X}$ and $\mathbf{Y}$ given the value of a third variable set $\mathbf{Z}$. Note that CMI is also symmetrical, i.e., $I(\mathbf{X};\mathbf{Y}|\mathbf{Z}) = I(\mathbf{Y};\mathbf{X}|\mathbf{Z})$.

2.2. Bayes error rate and mutual information

The optimal salient condition of features often means the minimal classification error in the context of supervised learning. In the classifier-independent situation, minimal error generally requires maximal statistical dependence of the class C on the data distribution in the subspace $\mathbb{R}^m$. This scheme is called maximal dependence (Peng et al., 2005). In the context of information theory, the maximal dependence is generally characterized in terms of MI. Fano's inequality reveals that the Bayes error rate of predicting C from a feature set $\mathbf{S}$ is lower bounded by the following expression dependent on the MI (Fano, 1961):

$$\frac{H(C) - I(\mathbf{S};C) - 1}{log(|C|)} \leq P(g(\mathbf{F}) \neq C), \tag{1}$$

where $\mathbf{F}$ denotes the feature set and $\mathbf{S} \subset \mathbf{F}$, g denotes a given learning method, and the maximal $I(\mathbf{S};C)$ corresponds to the minimal bound that enables finding a well-performing classifier of which the predicting error reaches the bound. This is the theoretical foundation of $I(\mathbf{S};C)$ as the goal function of feature selection. The feature selection task is therefore to find a feature subset $\mathbf{S}$ that jointly has maximal dependence on class C. However, it is difficult to find the optimal feature subset satisfying $I(\mathbf{S};C)$ because of **NP**-hardness (Albrecht, 2006). In addition, it is intractable to obtain an accurate estimation for multivariate distribution $P(F_1, ..., F_{|\mathbf{S}|})$ due to sample insufficiency and considerable computational cost (this is so-called the curse of dimensionality). Heuristics such as the incremental greedy search is thus applied by most of the commonly used information theoretic feature selection methods to obtain near-optimal solutions, i.e., to find feature sequences of which the top-ranked features are salient for data representation (Vinh et al., 2016).

3. Dimensionality of feature correlations and the heuristics

Suppose $\mathbf{S}_{i-1}$ $(i \geq 2)$ is the feature subset obtained at the $i-1$-th step, then $\max I(\mathbf{S}_i; C)$ can be easily implemented using the following expansion (Cover and Thomas, 1991)

$$I(\mathbf{S}_i; C) = I(F_i; C) + I(\mathbf{S}_{i-1}; C) - I(\mathbf{S}_{i-1}; F_i) + I(\mathbf{S}_{i-1}; F_i|C), \tag{2}$$

where $\mathbf{S}_i = \mathbf{S}_{i-1} \cup \{F_i\}$ and $F_i \in \mathbf{F} \setminus \mathbf{S}_{i-1}$. Since F_i satisfying $\max I(\mathbf{S}_i; C)$ is irrelevant to $I(\mathbf{S}_{i-1}; C)$, Eq. (2) can be reformulated as

$$\begin{aligned} I(F_i; C|\mathbf{S}_{i-1}) &= I(\mathbf{S}_i; C) - I(\mathbf{S}_{i-1}; C) \\ &= \underbrace{I(F_i; C)}_{\textit{class-relevance}} - \underbrace{I(\mathbf{S}_{i-1}; F_i)}_{\textit{redundancy}} + \underbrace{I(\mathbf{S}_{i-1}; F_i|C)}_{\textit{complementarity}}, \end{aligned} \tag{3}$$

Eq. (3) implies equivalence of $\max I(\mathbf{S}_i; C)$ and $\max I(F_i; C|\mathbf{S}_{i-1})$ when using an incremental greedy search strategy. It dissects the intrinsic structure of CMI and decomposes it explicitly into three *meta correlations*, i.e., class-relevance ($I(F_i; C)$), redundancy ($I(\mathbf{S}_{i-1}; F_i)$), and complementarity ($I(\mathbf{S}_{i-1}; F_i|C)$; also called the feature correlation within classes (Brown et al., 2012)). The de-composition of CMI is necessary because it distinguishes feature inner correlations into "harmful" and "useful" ones whereas the latter is often overlooked in the extant feature selection literature. This is in accordance with the observations of Guyon et al. (2006) and Brown et al. (2012), who observed that "correlation does not imply redundancy". We note here that those meta correlations are of potential value as they link to the basic effects in the statistical models: (a) redundancy in essence captures multi-collinearity among variables as it implies variable correlation that impairs the effectiveness of the model; and (b) complementarity is essentially associated with the interaction effects and the "V"-structure in causal analysis and inference, as such effects and structure are necessary for accurate and interpretable modeling. Thus, the meta criteria shown in Eq. (3) are expected to select not only discriminative but also interpretable features.

However, measuring the meta correlations of redundancy and complementarity succeeds the estimation of multivariate distribution $P(F_1, ..., F_{|\mathbf{S}_{k-1}|})$, which is practically intractable due to the sample insufficiency and the considerable computational cost. To address this issue, pairwise correlations (as the approximation of the original correlations) are examined intuitively in almost all the popular information theoretic methods (Zhang et al., 2019). To further investigate the relationship between the pairwise approximation and the primal meta correlations, we show another expansion of $I(\mathbf{S}_i; C)$ consisting of pairwise feature correlations and a metric named interaction information (Brown, 2009), as follows:

$$I(\mathbf{S}_i; C) = \sum_{F \in \mathbf{S}_i} I(F; C) - \sum_{\substack{F_1, F_2 \in \mathbf{S}_i \\ F_1 \neq F_2}} I(F_1; F_2) + \sum_{\substack{F_1, F_2 \in \mathbf{S}_i \\ F_1 \neq F_2}} I(F_1; F_2|C) + \mathcal{I}_{\geq 3}, \tag{4}$$

where $\mathcal{I}_{\geq 3}$ denotes the higher-order interaction information among more than three features and can take negative values [1]. Similar to Eq. (3), the following equations can be obtained from Eq. (4):

$$
\begin{aligned}
I(F_i; C|\mathbf{S}_{i-1}) &= I(\mathbf{S}_i; C) - I(\mathbf{S}_{i-1}; C) \\
&= \begin{pmatrix} 1 \\ -1 \\ 1 \\ 1 \end{pmatrix}^T \cdot \begin{pmatrix} \Delta \sum\limits_{F \in \mathbf{S}_{i-1} \cup \{F_i\}} I(F; C) \\ \Delta \sum_{\substack{F_1, F_2 \in \mathbf{S}_{i-1} \cup \{F_i\} \\ F_1 \neq F_2}} I(F_1; F_2) \\ \Delta \sum_{\substack{F_1, F_2 \in \mathbf{S}_{i-1} \cup \{F_i\} \\ F_1 \neq F_2}} I(F_1; F_2|C) \\ \Delta\, \mathcal{I}_{\geq 3} \end{pmatrix} \\
&= \begin{pmatrix} 1 \\ -1 \\ 1 \\ 1 \end{pmatrix}^T \cdot \begin{pmatrix} I(F_i; C) \\ \sum_{F \in \mathbf{S}_{i-1}} I(F_i; F) \\ \sum_{F \in \mathbf{S}_{i-1}} I(F_i; F|C) \\ \Delta\, \mathcal{I}_{\geq 3} \end{pmatrix}.
\end{aligned} \tag{5}
$$

With Eqs.(3) and (5), we obtain:

$$
\begin{pmatrix} 1 \\ -1 \\ 1 \\ 1 \end{pmatrix}^T \cdot \begin{pmatrix} I(F_i; C) \\ \sum_{F \in \mathbf{S}_{i-1}} I(F_i; F) \\ \sum_{F \in \mathbf{S}_{i-1}} I(F_i; F|C) \\ \Delta\, \mathcal{I}_{\geq 3} \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ 1 \\ 0 \end{pmatrix}^T \cdot \begin{pmatrix} I(F_i; C) \\ I(F_i; \mathbf{S}_{i-1}) \\ I(F_i; \mathbf{S}_{i-1}|C) \\ 0 \end{pmatrix}, \tag{6}
$$

Note that both $\Delta\, \mathcal{I}_{\geq 3} \geq 0$ and $\Delta\, \mathcal{I}_{\geq 3} < 0$ may hold true depending on specific features (Brown, 2009). Eq. (6) indicates that, the second and third entries of MI vector on the left-hand side of the equation can be viewed as the pairwise approximations of the meta correlations, i.e., redundancy and complementarity, on the right-hand side of the equation. Thus $\Delta\, \mathcal{I}_{\geq 3}$ can be illustrated as the *error term*, carrying information loss (when $\Delta\, \mathcal{I}_{\geq 3} > 0$) or information overload (when $\Delta\, \mathcal{I}_{\geq 3} < 0$). A space of potential criteria that relate to numerous existing relevance-redundancy heuristics can be derived from such a perspective by adjusting the weights of the summation of the pairwise MI terms, in such a way as to (hopefully) get close to the meta correlations, i.e.,

$$
J_{linearity}(F_i) = \begin{pmatrix} 1 \\ \alpha \\ \beta \end{pmatrix}^T \cdot \begin{pmatrix} I(F_i; C) \\ -\sum_{F \in \mathbf{S}_{k-1}} I(F_i; F) \\ \sum_{F \in \mathbf{S}_{k-1}} I(F_i; F|C) \end{pmatrix} \simeq \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}^T \cdot \begin{pmatrix} I(F_i; C) \\ -I(F_i; \mathbf{S}_{i-1}) \\ I(F_i; \mathbf{S}_{i-1}|C) \end{pmatrix} \tag{7}
$$

where $(1, \alpha, \beta)^T$ (denoted as $\boldsymbol{\delta}$) are the weighted vectors for linear combinations of the pairwise MI terms corresponding to the meta correlations, and $J_{linearity}(\cdot)$ denotes the feature criterion. Since the confusing error term $\Delta\, \mathcal{I}_{\geq 3}$ is no longer accounted for in Eq. (7), introduction of $\boldsymbol{\delta}$ can also be interpreted as

[1] For the detailed form of interaction information, please refer to Brown (2009).

compensation for discarding $\Delta\ \mathcal{I}_{\geq 3}$. Detailed information of the existing relevance-redundancy criteria that belong to this linear span (corresponding to all possible parameter combinations) are given in Tab. 1. Note that although these criteria aim to obtain approximate solutions, many of them are *complete heuristics* in that they cannot guarantee theoretical bound(s) of the meta correlations. We use the following theorem to formally illustrate this:

Theorem 1. *There always exist random variables* $X, Y, Z (X \neq Y, X \neq Z, Y \neq Z)$ *and* $X', Y', Z' (X' \neq Y', X' \neq Z', Y' \neq Z')$ *that satisfy the following relationships:*

(i) $I(X;Z) + I(Y;Z) \geq I(X,Y;Z)$

(ii) $I(X';Z') + I(Y';Z') < I(X',Y';Z')$

Proof of Theorem 1 is straightforward in terms of Eq. (2) and the fact that $I(X;Y|Z) \geq I(X;Y)$ and $I(X;Y|Z) < I(X;Y)$ may both hold true under different conditions (see Appendix A in Brown (2009)). Therefore, there is no unequal relationship between $\alpha \sum_{F \in \mathbf{S}_{k-1}} I(F_i;F)$ and $I(\mathbf{S}_{k-1};F)$ and that between $\beta \sum_{F \in \mathbf{S}_{k-1}} I(F_i;F|C)$ and $I(\mathbf{S}_{k-1};F|C)$ if α and β are predefined constant values. Those criteria may sometimes cause unexpected effects, for example, FOU (and MIFS) apply the (weighted) summation of pairwise MI as the approximations of the meta correlations, thus allowing redundancy and complementarity to swamp relevance when the size of selected feature set $\mathbf{S}$ grows. Other criteria like JMI (Yang and Moody, 1999), the second order approximation (SOA) (Guo and Nixon, 2009) and mRMR (Peng et al., 2005) (partially) guarantee loose lower bounds of the meta correlations, because they all employ the average pairwise form (i.e., $1/|\mathbf{S}| \sum_{F \in \mathbf{S}} I(F_i;F)$ and $1/|\mathbf{S}| \sum_{F \in \mathbf{S}} I(F_i;F|C)$) (which will be explained in the next section). This can be seen as the theoretical support to the observations in the literature that JMI and mRMR (especially JMI) often perform most effectively and can achieve best overall trade-off for accuracy/stability among the state-of-the-arts (Brown et al., 2012; Zhang et al., 2019). It inspired us to search for other potential tighter lower (or upper) bounds of the meta correlations.

However, known in the literature, the average pairwise form seems to be the only one bound within the linear span (Eq. (7)). A natural direction is to extend the linear span to the nonlinear one and find more possible bounds within it. Intuitively, the criteria with maximization or minimization operators, such as CMIM (Fleuret, 2004) and the joint mutual information maximization (JMIM, Bennasar et al., 2015), seem to be the typical points within the nonlinear span. In the next section, we will indicate that the *maximization* operator can guarantee tighter lower bounds than the *average* operator, and will then propose an effective information theoretic feature selection method based on such lower bounds.

Table 1: Some representative information theoretic feature selection criteria and their corresponding parameter vectors.

Criterion	$J(F)$	δ	Bounds	Source
MIM	$I(F;C)$	$(1,0,0)$	No	Lewis (1992)
mRMR	$I(F;C)-\frac{1}{\lvert\mathbf{S}\rvert}\sum_{F'\in\mathbf{S}} I(F;F')$	$\left(1,\frac{1}{\lvert\mathbf{S}\rvert},0\right)$	Partial	Peng et al. (2005)
MIFS	$I(F;C)-\beta\sum_{F'\in\mathbf{S}} I(F;F'),\quad \beta\in(0,1]$	$(1,\beta,0)$	No	Battiti (1994)
JMI (SOA)	$\sum_{F'\in\mathbf{S}} I(F,F';C)$	$\left(1,\frac{1}{\lvert\mathbf{S}\rvert},\frac{1}{\lvert\mathbf{S}\rvert}\right)$	Yes	Yang and Moody (1999); Guo and Nixon (2009)
FOU	$I(F;C)-\sum_{F'\in\mathbf{S}} I(F;F')+\sum_{F'\in\mathbf{S}} I(F;F'\lvert C)$	$(1,1,1)$	No	Meyer et al. (2008); Brown et al. (2012)
MRI †	$\sum_{F'\in\mathbf{S}}\left[I(F;C\lvert F')+I(F';C\lvert F)\right]$	$\left(1,\frac{2}{1+\lvert\mathbf{S}\rvert},\frac{2}{1+\lvert\mathbf{S}\rvert}\right)$	No	Wang et al. (2017)

† Intermediate steps for the re-written form of MRI are given in Appendix A.1.

4. Feature selection with lower bounds of feature inner correlations

4.1. Lower bounds of inner correlations of features

Before introducing the lower bounds, we first show the monotonicity of MI with respect to the scale of the feature set in Property 1.

Property 1. *Given a feature set* $\mathbf{S}$ *and a feature* F $(F\notin\mathbf{S})$*, for* $\forall\widetilde{\mathbf{S}}\subseteq\mathbf{S}$*, the following relationship holds:*

$$I(F;\widetilde{\mathbf{S}})\leq I(F;\mathbf{S}).$$

Property 1 can be easily proved by applying the chain rule of MI. It shows that the MI between a feature and a feature set will increase when the size of the set increases. We can hereby deduce that the criteria of some well-performed methods, e.g., mRMR, contain the lower bound of redundancy, and others, e.g., JMI and SOA, contain the lower bounds of redundancy and complementarity. We show this in Proposition 1.

Proposition 1. *Let* $\mathbf{F}$ *be the full set of features of dataset* D*,* $\mathbf{S}\subseteq\mathbf{F}$*, and* C *be the set of class labels of* D*. Denote the meta correlations* $I(F;\mathbf{S})$ *as* OPT_R *and* $I(F;\mathbf{S}|C)$ *as* OPT_C*. Then, for* $\forall F\in\mathbf{F}\setminus\mathbf{S}$*, the following relationships hold:*

(i) $\frac{1}{\lvert\mathbf{S}\rvert}\sum_{F_\mathbf{S}\in\mathbf{S}} I(F;F_\mathbf{S})\leq OPT_R,$

(ii) $\frac{1}{\lvert\mathbf{S}\rvert}\sum_{F_\mathbf{S}\in\mathbf{S}} I(F;F_\mathbf{S}|C)\leq OPT_C.$

Proposition 1 is straightforward given Property 1. Proposition 1 indicates that all the criteria that include $\frac{1}{\lvert\mathbf{S}\rvert}\sum_{F_\mathbf{S}\in\mathbf{S}} I(F;F_\mathbf{S})$ ($\frac{1}{\lvert\mathbf{S}\rvert}\sum_{F_\mathbf{S}\in\mathbf{S}} I(F;F_\mathbf{S}|C)$) can guarantee the lower bound(s) of redundancy (complementarity). Recall that some of the criteria, such as MIFS, FOU, and MRI, cannot guarantee any bounds of

redundancy or complementarity because there is no relationship between their pairwise approximate terms and the corresponding meta correlations.

In addition, the following question still remains: are there any closer bounds of redundancy and complementarity? We find better lower bounds with surprisingly simple forms from the nonlinear span that can be obtained even faster than the average pairwise forms shown in Proposition 1. We show them in Proposition 2.

Proposition 2. *Let* $\mathbf{F}$ *be the feature set of dataset* D, $\mathbf{S} \subseteq \mathbf{F}$, *and* C *be the set of class labels of* D. *Denote the meta correlations* $I(F;\mathbf{S})$ *as* OPT_R *and* $I(F;\mathbf{S}|C)$ *as* OPT_C. *Then, for* $\forall F \in \mathbf{F} \setminus \mathbf{S}$, *the following relationships hold:*

(i) $\frac{1}{|\mathbf{S}|}\sum_{F_\mathbf{S}\in\mathbf{S}} I(F;F_\mathbf{S}) \leq \max_{F_\mathbf{S}\in\mathbf{S}} I(F;F_\mathbf{S}) \leq OPT_R$,

(ii) $\frac{1}{|\mathbf{S}|}\sum_{F_\mathbf{S}\in\mathbf{S}} I(F;F_\mathbf{S}|C) \leq \max_{F_\mathbf{S}\in\mathbf{S}} I(F;F_\mathbf{S}|C) \leq OPT_C$.

Proposition 2 indicates that, the maximal pairwise forms are also the lower bounds of the meta correlations (hereafter *max-bounds*). More importantly, they are closer to OPT_R and OPT_C than the average pairwise forms (hereafter *avg-bounds*), respectively. Given the current selected feature set $\mathbf{S}$, the proposed feature evaluation criterion called the *lower bounds of redundancy and complementarity* (LBRC) can be presented as

$$J_{LBRC}(F) = I(F;C) - \max_{F_\mathbf{S}\in\mathbf{S}} I(F;F_\mathbf{S}) + \max_{F_\mathbf{S}\in\mathbf{S}} I(F;F_\mathbf{S}|C) \tag{8}$$

The proposed lower bounds aim at reducing the gaps between the global redundancy/complementarity (i.e., the meta correlations) and their pairwise approximations, respectively. When $F' = \arg\max J_{LBRC}(F)$ is added into $\mathbf{S}$, it is expected that $\max I(\mathbf{S};C)$ (equivalent to $\max I(F;C|\mathbf{S}\setminus\{F\})$ in terms of the chain rule of MI) can be obtained. With Eq. (6), the gap between the global redundancy and complementarity and the corresponding pairwise approximations, i.e., $\Delta\mathcal{I}_{\geq 3}$, can be derived as

$$\begin{aligned}\Delta\mathcal{I}_{\geq 3} \quad = \quad & \underbrace{\left(I(F;\mathbf{S}) - \sum_{F_\mathbf{S}\in\mathbf{S}} I(F_\mathbf{S};F)\right)}_{\text{redundancy gap}} \\ & + \underbrace{\left(\sum_{F_\mathbf{S}\in\mathbf{S}} I(F_\mathbf{S};F|C) - I(F;\mathbf{S}|C)\right)}_{\text{complementarity gap}}.\end{aligned} \tag{9}$$

From this perspective, using the proposed max-bounds will get a smaller *redundancy gap* and a smaller *complementarity gap* than the avg-bounds, and thus can not only be expected to achieve a small $\Delta\mathcal{I}_{\geq 3}$, but also guarantee more accurate results where some of the meta-correlations play an important role, e.g., V-structure identification in causal structure discovery (Judea, 2000; Ling et al., 2020) where the complementarity term $I(F;\mathbf{S}|C)$ may serve as the critical criterion.

4.2. Correlations with JMIM and CMIM

Recall that two successful criteria, namely, JMIM and CMIM, apply minimization operator as nonlinear combinations of pairwise MI terms. The original forms of JMIM and CMIM are

$$\begin{aligned} J_{JMIM}(F) &= \min_{F_{\mathbf{S}} \in \mathbf{S}} I(F, F_{\mathbf{S}}; C) \\ J_{CMIM}(F) &= \min_{F_{\mathbf{S}} \in \mathbf{S}} I(F; C | F_{\mathbf{S}}) \end{aligned} \tag{10}$$

Both can be seen as the approximations of $I(F, \mathbf{S}; C)$ and $I(F; C|\mathbf{S})$, respectively, from a pessimistic perspective. Brown et al. (2012) show that CMIM can be expanded to

$$J_{CMIM}(F) = I(F; C) + \min_{F_{\mathbf{S}} \in \mathbf{S}} \left[-I(F, F_{\mathbf{S}}) + I(F, F_{\mathbf{S}} | C) \right] \tag{11}$$

Similarly, the expansion of JMIM is given as (see Appendix A.2)

$$J_{JMIM}(F) = I(F; C) + \min_{F_{\mathbf{S}} \in \mathbf{S}} \left[I(F_{\mathbf{S}}; C) - I(F, F_{\mathbf{S}}) + I(F, F_{\mathbf{S}} | C) \right] \tag{12}$$

Although both expansions above contain the pairwise approximation terms of the meta correlations, they measure redundancy and complementarity in aggregate rather than individually. As a result, they do slightly obscure the strong link to our framework. In addition, both the original and the expanded versions of JMIM and CMIM seem unable to guarantee any theoretical bounds of either the meta correlations or their original goal functions. We find that the only difference in the expanded versions of JMIM and CMIM is that JMIM contains an additional term $I(F_{\mathbf{S}}; C)$. In this regard, JMIM is more robust than CMIM, as it not only considers the redundancy-complementarity term (i.e., the term $-I(F, F_{\mathbf{S}}) + I(F, F_{\mathbf{S}}|C)$ that CMIM focuses on), but also uses $I(F_{\mathbf{S}}; C)$ to confirm that $F \in \mathbf{S}$ is an intrinsically "pessimistic" representative of the currently selected feature set $\mathbf{S}$. However, the strategy of JMIM is not always preferable since a small value of $I(F_{\mathbf{S}}; C)$ will enhance the likelihood of encountering a false positive $F_{\mathbf{S}} \in \mathbf{S}$. In addition, it is noted that LBRC is more suitable for distributed parallel computing than CMIM and JMIM, because each of the terms in Eq. (8) can be conducted completely and independently. This facilitates LBRC a capable criterion for the feature selection tasks in the context of big data.

4.3. The proposed algorithm

The proposed criterion LBRC shown in Eq. (8) is straightforward to implement. We present in this section a fast implementation of LBRC shown in Algorithm 1.

In Algorithm 1, $Rel(F)$ records the MI between the candidate F and C (class-relevance), and $R_{\max}(F)$ records the maximum MI between the candidate F and $F_{\mathbf{S}_{k-1}}$ selected in the $k-1$-th step. Similarly, $C_{\max}(F)$ records the maximum CMI between the candidate F and $F_{\mathbf{S}_{k-1}}$ selected in the $k-1$-th step given C. Algorithm 1 works due to the following principle: since the values of $\max I(F; F_{\mathbf{S}_{k-1}})$ and $\max I(F; F_{\mathbf{S}_{k-1}}|C)$ for each candidate feature $F \in \mathbf{F} \setminus \mathbf{S}$ are known at the end of the k-th step, we only need to make a

Algorithm 1 A fast implementation of LBRC

1: **Input: D** /*dataset*/, **F** /*feature set*/, C /*class*/, δ /*expected # features to be selected*/
2: **Output: S** /*selected feature subset*/
3: **Initialize:** $\mathbf{S} \leftarrow \varnothing$, $Rel(F) \leftarrow 0$ for all $F \in \mathbf{F}$, $F_{new} \leftarrow \varnothing$, $R_{\max}(F) \leftarrow 0$ for all $F \in \mathbf{F}$, $C_{\max}(F) \leftarrow 0$ for all $F \in \mathbf{F}$, $Score(F) \leftarrow 0$ for all $F \in \mathbf{F}$, $\max\_Score \leftarrow 0$, $k \leftarrow 0$
4: **for all** $F \in \mathbf{F}$ **do**
5: $Rel(F) \leftarrow I(F;C)$
6: **if** $\max\_Score < Rel(F)$ **then**
7: $\max\_Score \leftarrow Rel(F)$, $F_{new} \leftarrow F$
8: **end if**
9: **end for**
10: $\mathbf{S} \leftarrow \mathbf{S} \cup \{F_{new}\}$, $\mathbf{F} \leftarrow \mathbf{F} \setminus \{F_{new}\}$, $k \leftarrow k+1$
11: **repeat**
12: $\max\_Score \leftarrow 0$
13: **for all** $F \in \mathbf{F}$ **do**
14: **if** $R_{\max}(F) < I(F;F_{new})$ **then**
15: $R_{\max}(F) \leftarrow I(F;F_{new})$
16: **end if**
17: **if** $C_{\max}(F) < I(F;F_{new}|C)$ **then**
18: $C_{\max}(F) \leftarrow I(F;F_{new}|C)$
19: **end if**
20: $Score(F) \leftarrow Rel(F) - R_{\max}(F) + C_{\max}(F)$
21: **if** $\max\_Score < Score(F)$ **then**
22: $\max\_Score \leftarrow Score(F)$, $F_{new} \leftarrow F$
23: **end if**
24: **end for**
25: $\mathbf{S} \leftarrow \mathbf{S} \cup \{F_{new}\}$, $\mathbf{F} \leftarrow \mathbf{F} - \{F_{new}\}$, $k \leftarrow k+1$
26: **until** $k \geq \delta$
27: **return S**

comparison between $\max I(F;F_{\mathbf{S}_{k-1}})$ and the MI value associated with the newly-added feature F_k, i.e., $I(F;F_k)$ (and that between $\max I(F;F_{\mathbf{S}_{k-1}}|C)$ and $I(F;F_k|C)$), and then choose the greater one as the value of $\max I(F;F_{\mathbf{S}_k})$ (and $\max I(F;F_{\mathbf{S}_k}|C)$), for each candidate feature $F \in \mathbf{F} \setminus \mathbf{S}$. This can be formed as

$$R^k_{\max}(F) = \max\left[R^{k-1}_{\max}(F), I(F;F_k)\right]$$

and

$$C^k_{\max}(F) = \max\left[C^{k-1}_{\max}(F), I(F;F_k|C)\right]$$

which correspond exactly to lines 14-16 and 17-19 in Algorithm 1. Based on this, the worst-case iteration complexity of Algorithm 1 is only $\mathrm{O}(\delta \cdot |\mathbf{F}|)$, which is very efficient and competent in the context of big data.

One key problem that remains here as well as in a series of existing feature selection works is to determine the number of the selected features δ. For most of the existing feature ranking methods, δ serves as a

hyperparameter and should be predefined for any specific learning tasks. However, the optimal number of selected features is often unknown. A larger δ will increase the likelihood of including redundant and irrelevant features, whereas a smaller δ may leave out relevant and complementary features. An acceptable approach is to grid search the number of selected features and select the number that corresponds to the best classification result, but the whole process has low efficiency and is sensitive to the specific classifier. How to optimally determine δ for a filter remains an open problem (Li et al., 2017).

5. Experimental study

5.1. Methods and datasets

In this section, we will empirically evaluate the performance of the proposed method. The most representative and well-performing feature selection methods using avg-bounds, namely, mRMR and JMI, are selected as the benchmark methods. Besides, the typical methods pointed within the linear span shown in Eq. (7), namely, MIM and FOU, the representative and relative methods JMIM, CMIM, and one classical instance-based method called ReliefF (Robnik-Sikonja and Kononenko, 2003), are used for comparison with the proposed method. Weka (Waikato Environment for Knowledge Analysis) (Witten and Frank, 2000) is chosen as the experimental platform. Since MIM and ReliefF have already been integrated into Weka, we directly call them in the data preprocessing panel to select features and generate the corresponding datasets. FOU, mRMR, JMI, and the proposed LBRC are implemented in Java with Weka interfaces. In our experiments, 5 neighbors and 30 instances are chosen for ReliefF, as suggested by Robnik-Sikonja and Kononenko (2003). All the experiments are conducted on a personal computer with a 2.60 GHz CPU, 8 GB of RAM and Windows 7.

To validate the performance of the proposed method, a total of twenty two public datasets with a wide range of dimensionalities are used in the experiments. We first conduct the comparison on eight UCI datasets. Then, two groups of well-known datasets, i.e., eight image datasets and six microarray datasets, are used to extend the experiments. Detailed information of these datasets is presented in Tab. 2. For the continuous and mixed datasets, a frequently used discretization method based on minimum description length (MDL) (Fayyad and Irani, 1993) is employed to discretize continuous features before feature selection.

5.2. Experimental procedure

In the experiments, four of the most frequently used classifiers, namely, C4.5 decision tree (Quinlan, 1993), k-nearest neighbor (kNN) (Aha and Kibler, 1991), naïve Bayes classifier (NBC) (Witten and Frank, 2000), and support vector machine (SVM) (Cristianini and Shawe-Taylor, 2000), are used to generate classification results over the datasets preprocessed by different feature selection methods. Following He et al. (2011) and

Table 2: Description of the selected datasets in our experiments.

Category	#	Name	# samples	# features	Type	# classes	CV
UCI	1	dermatology	366	34	mixed	6	10-fold
	2	spambase	4601	57	mixed	2	10-fold
	3	splice	3190	60	mixed	3	10-fold
	4	synthetic	600	60	mixed	6	10-fold
	5	optdigits	5620	63	nominal	10	10-fold
	6	isolet5	1559	617	mixed	26	10-fold
	7	gisette	6000	5000	mixed	2	10-fold
	8	nci9_s3	60	9712	nominal	9	Leave-One-Out
Image	9	mfeat-factors	2000	216	nominal	10	10-fold
	10	mfeat-karhunen	2000	64	numeric	10	10-fold
	11	mfeat-pixel	2000	240	nominal	10	10-fold
	12	mfeat-zernike	2000	47	mixed	10	10-fold
	13	PIE10P	210	2420	nominal	10	Leave-One-Out
	14	AR10P	130	2400	nominal	10	Leave-One-Out
	15	PIX10P	100	10000	nominal	10	Leave-One-Out
	16	ORL10P	100	10304	nominal	10	Leave-One-Out
Microarray	17	SRBCT	83	2308	mixed	4	Leave-One-Out
	18	14 Tumors	308	15009	mixed	26	Leave-One-Out
	19	lung cancer – Michigan	96	7129	mixed	2	Leave-One-Out
	20	ovarian PBSII	253	15154	mixed	2	Leave-One-Out
	21	colon tumor	62	2000	numeric	2	Leave-One-Out
	22	lymphoma	66	4026	mixed	3	Leave-One-Out

Kundu and Mitra (2017), we set $k = 1$ for kNN and apply the radial basis function (RBF) kernel for SVM, which are also default parameter settings for kNN and SVM in Weka.

First, we show the classification error rate of the four classifiers on the datasets represented with the top δ selected features for each feature selection method, where δ is the desired number of selected features specified as $\delta = [1, 2, ...t]$. The maximal acceptable size t is set to be $\min\{60, |\mathbf{F}|\}$ for the datasets containing less than 100 features, and it is set to $\min\{200, |\mathbf{F}|\}$ for the datasets containing more than 100 features. To obtain stable results, Cross-Validation (CV) is applied as a trade-off between stability and efficiency. Specifically, we apply 10-fold CV for all UCI datasets except nci9_s3 and some of image datasets ("mfeat" series), and apply Leave-One-Out CV for nci9_s3 and the rest of image datasets including PIE10P, AR10P, PIX10P, and ORL10P, and all microarray datasets to prevent data fragmentation. Given the aim of examining in detail the effectiveness of the proposed max-bounds in detail, we choose mRMR and JMI which apply avg-bounds as the comparison methods, and MIM and FOU (which are bound-free but belong to the linear span as mentioned before) and ReliefF (which is a typical instance-based feature selection method) as the benchmarks. For a clearer presentation, we do not include the results of JMIM and CMIM in this segment.

Note that the nature of the learning process of each classifier is different. Since we are interested in the quality of selected features rather than the performance of any specific classifiers, the average results of the four classifiers w.r.t. the number of selected features are reported to evaluate the overall quality of the selected features.

To ensure the robustness of the our evaluation, following Herman et al. (2013) and Wang et al. (2015), we perform additional experiments by statistically comparing the classification results corresponding to LBRC and any other method (i.e., mRMR, JMI, MIM, FOU, CMIM, JMIM) except ReliefF [2] with the same number of the selected features. For the datasets with less than 200 features, the top 10, 20, and 30 features are applied to obtain the classification results; and for the datasets with more than 200 (inclusive) features, the top 60, 80, and 100 features are applied [3]. In this segment, CV is repeated five times with different random seeds to generate sufficient classification error samples for the statistical test. As suggested by Demšar (2006), the Wilcoxon rank-sum test with a significance level of 0.05 is applied to determine the statistical significance of the differences between the results of LBRC and any other selected method.

5.3. Results and discussion

Figs. 1–3 show the average error rates of the four classifiers (C4.5, kNN, NBC, and SVM) w.r.t. the number of selected features on the selected datasets, where the number of the selected features is depicted on the X axis and the average classification error rate is presented on the Y axis. The results shown in Figs. 1–3 illustrate the comparableness and the effectiveness of LBRC on the whole. Specifically, LBRC significantly outperforms MIM, ReliefF, mRMR, DISR, FOU, and JMI in the majority of cases, particularly on the UCI datasets like dermatology (Fig. 1(a)), synthetic control (Fig. 1(c)), isolet5 (Fig. 1(f)), and gisette (Fig. 1(g)), the image datasets like mfeat-zernike (Fig. 2(d)), PIE10P (Fig. 2(e)), and AR10P (Fig. 2(f)), and the microarray datasets like ovarian PBSII (Fig. 3(d)), colon tumor (Fig. 3(e)), 14 tumors (Fig. 3(b)), and SRBCT (Fig. 3(a)).

Tabs. 3–5 summarize the results of kNN[4] on the selected UCI, image, and microarray datasets with the top 10, 20, and 30 (some are 60, 80, and 100) selected features, respectively.

For each dataset, the Wilcoxon rank-sum test is conducted to evaluate the statistical significance of the difference between two sequences of the samples of classification results, i.e., the sequence of the result

[2] ReliefF does not belong to the linear span shown in Eq. (7) and performs significantly inferiorly than other selected methods (see Figs. 1–3) and thus is omitted in statistical tests.

[3] We note here that in contrast, some of related studies, e.g., Blum and Langley (1997) and Song et al. (2013), choose to report the highest classification accuracy as well the corresponding number of selected features. This is appropriate for feature subset selection methods, i.e., the methods that can determine the selected feature subset with a certain number of features, but is unfair for feature ranking methods which aim to rank features rather than generate an optimal feature subset.

[4] The statistical results of C4.5, NBC, and SVM are similar to those of kNN and thus are omitted in this paper.

Table 3: Results of the classification error rate for kNN and the Wilcoxon test on UCI datasets with the top 10, 20, and 30 (some are 60, 80, and 100) selected features.

	#	#	LBRC	MIM		mRMR		JMI		FOU		CMIM		JMIM	
	dat.	fea.	Err	Err	p-val	Err	p-val	Err	p-val	Err	p-val	Err	p-val	Err	p-val
		10	**4.16**	23.09	0.000 ∘	15.53	0.000 ∘	15.53	0.000 ∘	14.82	0.000 ∘	5.86	0.006 ∘	5.95	0.002 ∘
	1	20	4.98	5.41	0.513	**2.62**	0.003 •	5.41	0.513	7.34	0.004 ∘	**4.98**	1.000	5.63	0.282
		30	4.43	4.16	0.812	4.16	0.812	3.95	0.384	4.98	0.391	4.76	0.630	**3.29**	0.191
		10	**7.71**	8.30	0.040 ∘	7.79	0.704	8.06	0.151	10.16	0.000 ∘	**7.71**	1.000	8.09	0.192
	2	20	**6.66**	7.83	0.000 ∘	7.24	0.020 ∘	7.47	0.000 ∘	9.46	0.000 ∘	7.18	0.037 ∘	7.94	0.000 ∘
		30	**6.40**	7.69	0.000 ∘	6.56	0.631	7.31	0.000 ∘	8.95	0.000 ∘	6.98	0.050 ∘	6.93	0.029 ∘
		10	13.19	12.90	0.447	13.19	1.000	12.90	0.447	14.23	0.017 ∘	12.90	0.447	**12.51**	0.055
	3	20	18.04	18.04	1.000	**17.32**	0.073	18.04	1.000	20.52	0.000 ∘	**17.32**	0.073	18.04	1.000
		30	**20.49**	22.28	0.000 ∘	21.93	0.001 ∘	22.28	0.000 ∘	23.42	0.000 ∘	21.72	0.001 ∘	22.28	0.000 ∘
		10	**5.07**	22.97	0.000 ∘	15.17	0.000 ∘	14.67	0.000 ∘	9.83	0.000 ∘	5.96	0.255	17.63	0.000 ∘
	4	20	**3.20**	18.10	0.000 ∘	11.70	0.000 ∘	11.60	0.000 ∘	8.23	0.000 ∘	4.09	0.047	13.77	0.000 ∘
kNN		30	**1.27**	10.63	0.000 ∘	8.53	0.000 ∘	5.63	0.000 ∘	8.80	0.000 ∘	3.12	0.000 ∘	8.50	0.000 ∘
		10	**13.57**	16.17	0.000 ∘	14.81	0.000 ∘	15.00	0.000 ∘	16.13	0.000 ∘	14.28	0.010 ∘	14.69	0.000 ∘
	5	20	8.38	9.58	0.000 ∘	8.75	0.098	9.56	0.000 ∘	10.40	0.000 ∘	**8.25**	0.621	8.40	0.764
		30	6.20	6.91	0.001 ∘	6.56	0.064	6.56	0.064	7.59	0.000 ∘	6.62	0.045 ∘	**6.02**	0.451
		60	26.80	35.84	0.000 ∘	28.58	0.008 ∘	31.62	0.000 ∘	28.42	0.014 ∘	**26.12**	0.061	33.01	0.000 ∘
	6	80	**22.47**	31.30	0.000 ∘	25.12	0.000 ∘	27.22	0.000 ∘	26.93	0.000 ∘	22.82	0.683	29.61	0.000 ∘
		100	**21.53**	27.70	0.000 ∘	21.55	0.959	22.33	0.198	26.03	0.000 ∘	21.92	0.701	25.73	0.000 ∘
		60	**5.62**	7.71	0.000 ∘	5.99	0.030 ∘	6.47	0.000 ∘	7.31	0.000 ∘	6.09	0.047 ∘	7.47	0.000 ∘
	7	80	**5.02**	7.12	0.000 ∘	5.51	0.007 ∘	6.26	0.000 ∘	7.19	0.000 ∘	5.69	0.000 ∘	6.46	0.000 ∘
		100	5.17	6.12	0.000 ∘	**5.04**	0.516	5.50	0.051	6.98	0.000 ∘	5.42	0.223	5.94	0.000 ∘
		60	**8.33**	20.00	0.000 ∘	15.00	0.000 ∘	16.67	0.000 ∘	43.33	0.000 ∘	15.00	0.000 ∘	11.67	0.002 ∘
	8	80	**8.33**	23.33	0.000 ∘	13.33	0.000 ∘	16.67	0.000 ∘	45.00	0.000 ∘	11.67	0.002 ∘	**8.33**	1.000
		100	**11.67**	23.33	0.000 ∘	15.00	0.008 ∘	13.33	0.031 ∘	46.67	0.000 ∘	16.67	0.000 ∘	**11.67**	1.000
	Avg.		**9.95**	15.69		12.37		12.92		17.20		10.96		12.48	
	L/W/T			20/0/4		14/1/9		16/0/8		23/0/1		12/0/12		14/0/9	

∘ statistical degradation at significance level of 0.05.

• statistical improvement at significance level of 0.05.

Table 4: Results of the classification error rate for kNN and the Wilcoxon test on image datasets with the top 10, 20, and 30 (some are 60, 80, and 100) selected features.

	#	#	LBRC	MIM		mRMR		JMI		FOU		CMIM		JMIM	
	dat.	fea.	Err	Err	p-val	Err	p-val	Err	p-val	Err	p-val	Err	p-val	Err	p-val
		60	**5.83**	9.40	0.000 ◦	6.24	0.063	7.23	0.000 ◦	5.98	0.525	**4.98**	0.010 •	5.92	0.510
	9	80	5.73	6.52	0.008 ◦	**5.10**	0.039 •	5.80	0.662	**5.63**	0.612	**4.67**	0.001 •	6.03	0.311
		100	5.14	5.73	0.052	**4.92**	0.598	6.13	0.002 ◦	**4.72**	0.225	**4.29**	0.005 •	5.11	0.961
		10	13.41	**13.37**	0.792	**13.37**	0.792	**13.37**	0.792	13.71	0.493	13.41	1.000	**13.37**	0.792
	10	20	8.94	9.22	0.323	8.88	0.989	9.11	0.624	**8.34**	0.087	9.32	0.208	8.88	0.989
		30	**7.86**	8.06	0.484	8.06	0.484	7.91	0.747	8.46	0.123	8.02	0.371	8.18	0.224
		60	**6.49**	44.14	0.000 ◦	20.68	0.000 ◦	26.01	0.000 ◦	26.34	0.000 ◦	7.39	0.004 ◦	7.74	0.000 ◦
	11	80	5.39	35.97	0.000 ◦	14.26	0.000 ◦	18.81	0.000 ◦	22.47	0.000 ◦	**5.17**	0.555	6.88	0.000 ◦
		100	5.55	27.86	0.000 ◦	13.12	0.000 ◦	15.26	0.000 ◦	23.47	0.000 ◦	**5.03**	0.115	5.70	0.739
		10	**36.22**	43.19	0.000 ◦	38.58	0.000 ◦	37.89	0.001 ◦	36.79	0.163	**36.22**	1.000	36.78	0.233
	12	20	29.38	36.37	0.000 ◦	31.00	0.003 ◦	35.08	0.000 ◦	35.25	0.000 ◦	**28.67**	0.121	32.03	0.000 ◦
kNN		30	**26.49**	30.75	0.000 ◦	29.82	0.000 ◦	30.23	0.000 ◦	31.78	0.000 ◦	27.09	0.225	29.36	0.000 ◦
		60	0.48	3.81	0.000 ◦	2.38	0.000 ◦	1.90	0.003 ◦	**0.00**	0.025 •	0.48	1.000	2.38	0.000 ◦
	13	80	0.48	2.38	0.000 ◦	1.43	0.025 ◦	1.90	0.003 ◦	**0.00**	0.025 •	0.48	1.000	1.43	0.025 ◦
		100	**0.00**	1.90	0.000 ◦	1.43	0.000 ◦	1.90	0.000 ◦	0.48	0.025 ◦	0.48	0.025 ◦	0.95	0.002 ◦
		60	**3.08**	9.23	0.000 ◦	4.62	0.149	6.92	0.001 ◦	9.23	0.000 ◦	4.62	0.149	7.69	0.000 ◦
	14	80	**2.31**	7.69	0.000 ◦	4.62	0.023 ◦	6.92	0.000 ◦	8.46	0.000 ◦	3.85	0.108	6.15	0.001 ◦
		100	**2.31**	5.38	0.004 ◦	3.08	0.392	4.62	0.023 ◦	7.69	0.000 ◦	3.08	0.392	6.15	0.001 ◦
		60	**0.00**	**0.00**	1.000	3.00	0.000 ◦	1.00	0.025 ◦	7.00	0.000 ◦	**0.00**	1.000	**0.00**	1.000
	15	80	1.00	3.00	0.024 ◦	**0.00**	0.025 •	**0.00**	0.025 •	6.00	0.000 ◦	1.00	1.000	**0.00**	0.025 •
		100	1.00	3.00	0.024 ◦	**0.00**	0.025 •	**0.00**	0.025 •	5.00	0.000 ◦	1.00	1.000	**0.00**	0.025 •
		60	**0.00**	4.00	0.000 ◦	**0.00**	1.000	**0.00**	1.000	4.00	0.000 ◦	**0.00**	1.000	**0.00**	1.000
	16	80	**0.00**	3.00	0.000 ◦	**0.00**	1.000	**0.00**	1.000	6.00	0.000 ◦	**0.00**	1.000	**0.00**	1.000
		100	**0.00**	**0.00**	1.000	**0.00**	1.000	**0.00**	1.000	3.00	0.000 ◦	**0.00**	1.000	**0.00**	1.000
	Avg.		**6.96**	13.08		8.94		9.91		11.66		7.05		7.95	
	L/W/T			20/0/4		9/5/10		14/4/6		16/0/8		13/4/7		13/0/11	

◦ statistical degradation at significance level of 0.05.

• statistical improvement at significance level of 0.05.

Table 5: Results of the classification error rate for kNN and the Wilcoxon test on microarray datasets with the top 60, 80, and 100 selected features.

	#	#	LBRC	MIM		mRMR		JMI		FOU		CMIM		JMIM	
	dat.	fea.	Err	Err	p-val	Err	p-val	Err	p-val	Err	p-val	Err	p-val	Err	p-val
kNN	17	60	**0.00**	**0.00**	1.000	**0.00**	1.000	**0.00**	1.000	**0.00**	1.000	**0.00**	1.000	**0.00**	1.000
		2	**0.00**	**0.00**	1.000	**0.00**	1.000	**0.00**	1.000	1.20	0.025 ∘	**0.00**	1.000	**0.00**	1.000
		3	**0.00**	**0.00**	1.000	**0.00**	1.000	**0.00**	1.000	**0.00**	1.000	**0.00**	1.000	**0.00**	1.000
	18	60	32.79	41.88	0.000 ∘	37.99	0.003 ∘	38.96	0.000 ∘	**30.52**	0.175	30.84	0.246	39.29	0.000 ∘
		80	27.60	38.64	0.000 ∘	37.34	0.000 ∘	40.91	0.000 ∘	31.17	0.030 ∘	**23.05**	0.004 •	38.31	0.000 ∘
		100	**24.03**	42.21	0.000 ∘	32.79	0.000 ∘	36.69	0.000 ∘	31.49	0.000 ∘	24.35	0.833	32.79	0.000 ∘
	19	60	**0.00**	**0.00**	1.000	**0.00**	1.000	**0.00**	1.000	**0.00**	1.000	1.04	0.025 ∘	**0.00**	1.000
		80	**0.00**	**0.00**	1.000	**0.00**	1.000	**0.00**	1.000	**0.00**	1.000	1.04	0.025 ∘	**0.00**	1.000
		100	**0.00**	**0.00**	1.000	**0.00**	1.000	**0.00**	1.000	**0.00**	1.000	**0.00**	1.000	**0.00**	1.000
	20	60	**0.00**	**0.00**	1.000	0.40	0.025 ∘	**0.00**	1.000	0.40	0.025 ∘	**0.00**	1.000	**0.00**	1.000
		80	**0.00**	0.40	0.025 ∘	0.40	0.025 ∘	0.40	0.025 ∘	0.79	0.002 ∘	**0.00**	1.000	**0.00**	1.000
		100	0.40	0.79	0.196	0.40	1.000	0.79	0.196	0.79	0.196	**0.00**	0.025 •	0.40	1.000
	21	60	12.90	16.13	0.255	12.90	1.000	**11.29**	0.539	12.90	1.000	**11.29**	0.539	12.90	1.000
		80	16.13	**11.29**	0.080	12.90	0.255	14.52	0.578	14.52	0.578	12.90	0.255	**11.29**	0.080
		100	14.52	**12.90**	0.560	**12.90**	0.560	14.52	1.000	14.52	1.000	16.13	0.578	**12.90**	0.560
	22	60	**0.00**	1.52	0.025 ∘	**0.00**	1.000	**0.00**	1.000	9.09	0.000 ∘	1.52	0.025 ∘	**0.00**	1.000
		80	**0.00**	**0.00**	1.000	**0.00**	1.000	**0.00**	1.000	9.09	0.000 ∘	1.52	0.025 ∘	1.52	0.025 ∘
		100	**0.00**	**0.00**	1.000	**0.00**	1.000	**0.00**	1.000	6.06	0.000 ∘	1.52	0.025 ∘	**0.00**	1.000
	Avg.		**6.86**	9.13		8.17		8.65		9.07		7.70		12.86	
	L/W/T			6/0/12		5/0/13		4/0/14		8/0/10		5/1/12		7/0/11	

∘ statistical degradation at significance level of 0.05.

• statistical improvement at significance level of 0.05.

samples corresponding to LBRC and that corresponding to any other feature selection method. In Tabs. 3–5, *# dat.* records the number of datasets, and *# fea.* shows the number of selected features. *Err* records the average classification error rate. *p-val* records the p-value associated with the Wilcoxon rank-sum test. The notation $\bullet/\circ$ is used to show that the average error rate corresponding to the current feature selection method is significantly lower/higher than that corresponding to the proposed method listed in the *LBRC* column. The bold value in each row indicates the best result among seven feature selection methods. The average error rates of all used datasets and the loss/win/tie (which records the times of significant degradation/significant improvement/no significant difference compared with the proposed method) are given in the last two rows, respectively. In can be concluded from the tables that LBRC performs significantly best in most of cases. The results of loss/win/tie consistently imply the superiority of LBRC.

On the basis of the experimental results, we arrive at the following observations:

(i) *Necessity of the meta correlations*: The inferior performance of MIM and ReliefF illustrates that neglecting redundancy and complementarity explicitly may lead to severely low quality of the selected features that may contain a considerable amount of redundancy.

(ii) *Necessity of the dynamic weights for the meta correlations*: FOU usually performs comparably at the very beginning, but relatively inferiorly to LBRC, mRMR, JMI, CMIM, and JMIM when the number of the selected features grows. This is possibly because the weighted vector of FOU, i.e., $\delta = (1, 1, 1)$ which corresponds to cumulative approximations of the meta correlations, gives rise to the situation where the class-relevance term is overwhelmed by the redundancy and the complementarity terms (i.e., overemphasis of redundancy and complementarity). We also notice an intersection point of the results of FOU and MIM in some cases (e.g., results on mfeat-zernike, ovarian PBSII, lymphoma, and colon tumor). This demonstrates the necessity of dynamically trading-off the approximations of the meta correlations, while implying an attractive potential in finding a dynamic balance empirically between the class-relevance and the feature inner correlations.

(iii) *Advantages of the bounded approximations of the meta correlations*: LBRC, mRMR, and JMI outperform MIM and FOU in general, indicating the effectiveness of the bounded approximations of the meta correlations. Features that satisfy the lower-bounded evaluation (both for the max-bounds and the avg-bounds) are expected to be more reliable than those obtained by complete heuristics.

(iv) *Limitations of the avg-bounds*: Experimental results indicate that there is no significant gap between the performance of mRMR and JMI, suggesting that the avg-bound of the complementarity has no significant impact on the quality of the selected features. This is probably because the gap between the complementarity and its avg-bound may interfere the results in some cases, which makes us believe that the avg-bound is not always a qualified lower bound.

(v) *Advantages of the proposed max-bounds*: It is evidenced that LBRC (using the max-bounds of redun-

dancy and complementarity) significantly outperforms mRMR (using the avg-bound for redundancy only) and JMI (using the avg-bound for redundancy and complementarity) in most cases (especially on the datasets of isolet5, ORL10P, 14 tumors, ovarian PBSII, colon tumor, lymphoma, lung cancer – Michigan). The key difference between LBRC and the benchmark methods (mRMR and JMI) is the consideration of max-bounds instead of avg-bounds as the approximations of the meta correlations. Methodologically, the max-bounds are closer to the meta correlations than the avg-bounds; Empirically, LBRC corresponds to better classification results than mRMR and JMI on the whole, indicating that tighter bounds of the meta correlations will lead to a better quality of the selected features.

(vi) *Effectiveness of separately approximating the meta correlations*: Statistical results from Tabs. 3–5 show that the performance of LBRC is superior to those of JMIM and CMIM in general. This is in consistent with the observations of Brown et al. (2012), implying that approximating the meta correlations separately rather than comprehensively may give rise to more robust results.

6. Conclusions & future work

Redundancy and complementarity are two important dimensions of feature inner correlations and attract considerable attention in big data analytics. Many information theoretic feature selection methods select features with heuristics that cannot guarantee theoretical bounds of those correlations. In this paper, we reformulate and analyze some representative information theoretic feature selection methods, including mRMR, JMI, and FOU, from the perspective of the lower bounds of feature inner correlations. Then, we introduce two lower bounds of the two essential dimensions of feature correlations, namely, redundancy and complementarity, and we theoretically verify that they are better than those applied in the existing information theoretic methods such as mRMR and JMI.

To evaluate the effectiveness of the proposed method LBRC, experiments are conducted with four popular classifiers (C4.5, NBC, kNN, and SVM) on twenty one publicly available real-world datasets. In the experiments, six representative feature selection methods, namely, MIM, mRMR, JMI, FOU, DISR, and ReliefF, are used for comparison with LBRC. The classification results illustrate the superiority of LBRC. To obtain reliable results, Wilcoxon rank-sum tests are conducted to statistically distinguish the performance of the selected method. According to the experimental results, LBRC significantly outperforms all compared feature selection methods on the whole, thus further verifying the superiority of the proposed lower bounds of feature inner correlations. It is worth noting that, despite the proposed max-bounds guarantee smaller gaps for the corresponding meta correlations than the avg-bounds, they themselves are not/do not form the bounds of the goal function of feature selection. For that matter, the proposed feature selection criterion is still heuristic. However, for some tasks like causal discovery that explicitly focus on individual meta correlations, the proposed lower bounds appear to be of more theoretical importance. For example, Yu et al. (2020)

utilize the pairwise avg-bound of $I(C;\mathbf{S})$ as the approximation to find the invariant sets of causal features, which seems less effective than using the max-bound. Possible future work includes using the proposed bounds for causal structure discovery and model interpretation and integrating the lower-bounded pairwise approximations into the discrete programming formulation. In addition, involving kernel methods for probability estimation on continuous features and finding closer bounds for k-wise ($k \geq 3$) inner correlations of features will also be considered in the future.

Acknowledgments

We would like to thank the editor and two anonymous referees for their constructive comments and suggestions. This study was supported in part by the National Natural Science Foundation of China under Grants 71702066 and 61703319, and in part by the Fundamental Research Funds for the Central Universities under Grant WUT: 2020IVA007. Zhijun Chen serves as the first corresponding author of this paper.

References

Aha, D., Kibler, D., 1991. Instance-based learning algorithms. Machine Learning 6, 37–66.

Albrecht, A. A., 2006. Stochastic local search for the feature set problem, with applications to microarray data. Applied Mathematics and Computation 183 (2), 1148–1164.

Antonov, A. V., Tetko, I. V., Mader, M. T., Budczies, J., Mewes, H. W., 2004. Optimization models for cancer classification: extracting gene interaction information from microarray expression data. Bioinformatics 20 (5), 644–652.

Aytug, H., 2015. Feature selection for support vector machines using generalized benders decomposition. European Journal of Operational Research 244 (1), 210–218.

Battiti, R., 1994. Using mi for selecting features in supervised neural net learning. IEEE Transactions on Neural Networks 5 (4), 537–550.

Bennasar, M., Hicks, Y., Setchi, R., 2015. Feature selection using joint mutual information maximisation. Expert Systems with Applications 42 (22), 8520–8532.

Bennasar, M., Setchi, R., Hicks, Y., 2013. Feature interaction maximisation. Pattern Recognition Letters 34 (14), 1630–1635.

Bertolazzi, P., Felici, G., Festa, P., Fiscon, G., Weitschek, E., 2016. Integer programming models for feature selection: New extensions and a randomized solution algorithm. European Journal of Operational Research 250 (2), 389–399.

Bertsimas, D., Pauphilet, J., Parys, B. V., 2020. Sparse classification: a scalable discrete optimization perspective.

Blum, A., Langley, P., 1997. Selection of relevant features and examples in machine learning. Artificial Intelligence 97, 245–271.

Boln-Canedo, V., Alonso-Betanzos, A., 2019. Ensembles for feature selection: A review and future trends. Information Fusion 52, 1–12.

Brown, G., 16–18 Apr 2009. A new perspective for information theoretic feature selection. In: van Dyk, D., Welling, M. (Eds.), Proceedings of the 12th International Conference on Artificial Intelligence and Statistics. Vol. 5 of Proceedings of Machine Learning Research. PMLR, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, pp. 49–56.
URL http://proceedings.mlr.press/v5/brown09a.html

Brown, G., Pocock, A., Zhao, M.-J., Luján, M., 2012. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. Journal of Machine Learning Research 13, 27–66.

Chang, X., Nie, F., Yang, Y., Huang, H., 2014. A convex formulation for semi-supervised multi-label feature selection. In: Proceedings of AAAI Conference on Artificial Intelligence. pp. 1171–1177.

Chen, J., Song, L., Wainwright, M., Jordan, M., 10–15 Jul 2018. Learning to explain: An information-theoretic perspective on model interpretation. In: Dy, J., Krause, A. (Eds.), Proceedings of the 35th International Conference on Machine Learning. Vol. 80. PMLR, Stockholmsmässan, Stockholm Sweden, pp. 883–892.
URL http://proceedings.mlr.press/v80/chen18j.html

Chen, Q., Chen, Y.-P. P., 2009. Discovery of structural and functional features in rna pseudoknots. IEEE Transaction on Knowledge and Data Engineering 21 (7), 974–984.

Chen, X., Xu, F., Ye, Y., 2010. Lower bound theory of nonzero entries in solutions of ℓ_2-ℓ_p minimization. SIAM Journal on Scientific Computing 32 (5), 2832–2852.

Cover, T. M., Thomas, J. A., 1991. Elements of Information Theory. Wiley, NewYork, NY, USA.

Cristianini, N., Shawe-Taylor, J., 2000. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press, Cambridge, UK.

Das, A. K., Pati, S. K., Ghosh, A., 2020. Relevant feature selection and ensemble classifier design using bi-objective genetic algorithm. Knowledge and Information Systems 62, 423–455.

Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7, 1–30.

Ding, C., Peng, H., 2003. Minimum redundancy feature selection from microarray gene expression data. In: Proceedings of the IEEE Computer Society Conference on Bioinformatics. CSB'03. IEEE Computer Society, Washington, DC, USA, pp. 523–528.

Fano, R., 1961. Transmission of Information: Statistical Theory of Communications. Wiley, New York, USA.

Fayyad, U. M., Irani, K. B., 1993. Multi-interval discretization of continuous valued attributes for classification learning. In: Proceedings of the 13th International Joint Conference on Artificial Intelligence. IJCAI'93. pp. 1022–1027.

Fleuret, F., 2004. Fast binary feature selection with conditional mutual information. Journal of Machine Learning Research 5, 1531–1555.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., Haussler, D., 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16 (10), 906–914.

Gao, W., Hu, L., Zhang, P., 2020. Feature redundancy term variation for mutual information-based feature selection. Applied Intelligence 50, 1272–1288.

Ghaddar, B., Naoum-Sawaya, J., 2018. High dimensional data classification and feature selection using support vector machines. European Journal of Operational Research 265 (3), 993–1004.

Guo, B., Nixon, M. S., 2009. Gait feature subset selection by mutual information. IEEE Transactions on Systems, Man and Cybernetics 39 (1), 36–46.

Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L., 2006. Feature Extraction: Foundations and Applications. Springer.

He, X., Ji, M., Zhang, C., Bao, H., 2011. A variance minimization criterion to feature selection using laplacian regularization. IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (10), 2013–2025.

Herman, G., Zhang, B., Wang, Y., Ye, G., Chen, F., 2013. Mutual information-based method for selecting informative feature sets. Pattern Recognition 46 (12), 3315–3327.

Judea, P., 2000. Causality: models, reasoning, and inference. Cambridge University Press, New York, NY, USA.

Kundu, P. P., Mitra, S., 2017. Feature selection through message passing. IEEE Transactions on Cybernetics 47 (12), 4356–4366.

Lewis, D. D., 1992. Feature selection and feature extraction for text categorization. In: Proceedings of the workshop on Speech and Natural Language. Association for Computational Linguistics Morristown, NJ, USA, pp. 212–217.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., Liu, H., 2017. Feature selection: A data perspective. ACM Computing Surveys 50 (6), 94:1–94:45.

Ling, Z., Yu, K., Wang, H., Li, L., Wu, X., 2020. Using feature selection for local causal structure learning. IEEE Transactions on Emerging Topics in Computational Intelligence, 1–11.

Liu, Y., Tang, F., Zeng, Z., 2015. Feature selection based on dependency margin. IEEE Transactions on Cybernetics 45 (6), 1209–1221.

López, J., Maldonado, S., 2019. Profit-based credit scoring based on robust optimization and feature selection. Information Sciences 500, 190–202.

Maldonado, S., Montoya, R., Weber, R., 2015. Advanced conjoint analysis using feature selection via support vector machines. European Journal of Operational Research 241, 564–574.

Maldonado, S., Pérez, J., Bravo, C., 2017. Cost-based feature selection for support vector machines: An application in credit scoring. European Journal of Operational Research 261 (2), 656–665.

Meyer, P. E., Schretter, C., Bontempi, G., 2008. Information-theoretic feature selection in microarray data using variable complementarity. IEEE Journal of Selected Topics in Signal Processing 2 (3), 261–274.

Peng, H., Long, F., Ding, C., 2005. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (8), 1226–1238.

Qu, G., Hariri, S., Yousif, M., 2005. A new dependency and correlation analysis for features. IEEE Transactions on Knowledge and Data Engineering 17 (9), 1199–1207.

Quinlan, R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA, USA.

Robnik-Sikonja, M., Kononenko, I., 2003. Theoretical and empirical analysis of relief and relieff. Machine Learning 53, 23–69.

Song, Q., Ni, J., Wang, G., 2013. A fast clustering-based feature subset selection algorithm for high dimensional data. IEEE Transactions on Knowledge and Data Engineering 25 (1), 1–14.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Ser. B 58, 267–288.

Vinh, N. X., Zhou, S., Chan, J., Bailey, J., 2016. Can high-order dependencies improve mutual information based feature selection? Pattern Recognition 53, 46–58.

Wang, D., Nie, F., Huang, H., 2015. Feature selection via global redundancy minimization. IEEE Transactions on Knowledge and Data Engineering 27 (10), 2743–2755.

Wang, J., Wei, J.-M., Yang, Z., Wang, S.-Q., 2017. Feature selection by maximizing independent classification information. IEEE Transactions on Knowledge and Data Engineering 29 (4), 828–841.

Witten, H. I., Frank, E., 2000. Data mining: practical machine learning tools and techniques with Java implementations. Morgan Kaufmann, San Francisco, CA, USA.

Won, D., Manzour, H., Chaovalitwongse, W., 2020. Convex optimization for group feature selection in networked data. INFORMS Journal on Computing 32 (1), 182–198.

Yang, H. H., Moody, J., 1999. Feature selection based on joint mutual information. In: Proceedings of International ICSC Symposium on Advances in Intelligent Data Analysis. pp. 22–25.

Yu, K., Liu, L., Li, J., Ding, W., Le, T. D., 2020. Multi-source causal feature selection. IEEE Transactions on Pattern Analysis and Machine Intelligence 42 (9), 2240–2256.

Yu, K., X. Xu, M. E., Kriegel, H.-P., 2003. Feature weighting and instance selection for collaborative filtering: An informationtheoretic approach. Knowledge and Information Systems 5 (2), 201–224.

Yu, L., Liu, H., 2004. Efficient feature selection via analysis of relevance and redundancy. Journal of Machine Learning Research 5, 1205–1224.

Zhang, Y., Yang, A., Xiong, C., Zhang, Z., 2014. Feature selection using data envelopment analysis. Knowledge-Based Systems 64, 70–80.

Zhang, Y., Zhang, Q., Chen, Z., Shang, J., Wei, H., 2019. Feature assessment and ranking for classification with nonlinear

sparse representation and approximate dependence analysis. Decision Support Systems 122, 113064.

Appendix A.

The following deductions make use of the information identity (Brown, 2009)

$$I(A;B|C) - I(A;B) = I(A;C|B) - I(A;C). \tag{A.1}$$

Appendix A.1. Intermediate steps for the re-written form of MRI

The max-relevance and max-independence (MRI) criterion (Wang et al., 2017) can be written as

$$\begin{aligned} J_{MRI}(F_i) &= I(F_i;C) + \sum_{F\in\mathbf{S}} ICI(C;F_i,F) \\ &= I(F_i;C) + \sum_{F\in\mathbf{S}} [I(F_i;C|F) + I(F;C|F_i)] \end{aligned} \tag{A.2}$$

With Eq. (A.1) we have

$$I(F_i;C|F) = I(F_i;C) - I(F_i;F) + I(F_i;F|C) \tag{A.3}$$

$$I(F;C|F_i) = I(F;C) - I(F_i;F) + I(F_i;F|C) \tag{A.4}$$

Then with Eqs. (A.3) and (A.4), Eq. (A.2) can be re-written as

$$I(F_i;C) + \sum_{F\in\mathbf{S}} [I(F_i;C) - I(F_i;F) + I(F_i;F|C) + I(F;C) - I(F_i;F) + I(F_i;F|C)] \tag{A.5}$$

The term $I(F;C)$ is constant with respect to the F_i argument that we focused and thus can be omitted. The criterion then reduces to

$$\begin{aligned} J_{MRI}(F_i) &= I(F_i;C) + \sum_{F\in\mathbf{S}} [I(F_i;C) - I(F_i;F) + I(F_i;F|C) - I(F_i;F) + I(F_i;F|C)] \\ &= (|\mathbf{S}|+1)I(F_i;C) + 2\sum_{F\in\mathbf{S}} [-I(F_i;F) + I(F_i;F|C)] \\ &\propto I(F_i;C) - \frac{2}{|\mathbf{S}|+1}\sum_{F\in\mathbf{S}} I(F_i;F) + \frac{2}{|\mathbf{S}|+1}\sum_{F\in\mathbf{S}} I(F_i;F|C). \end{aligned} \tag{A.6}$$

Note that the meta correlations are not bounded by $\frac{2}{|\mathbf{S}|+1}\sum_{F\in\mathbf{S}} I(F_i;F)$ and $\frac{2}{|\mathbf{S}|+1}\sum_{F\in\mathbf{S}} I(F_i;F|C)$, since $\frac{2}{|\mathbf{S}|+1}\sum_{F\in\mathbf{S}} I(F_i;F) \geq I(F_i;\mathbf{S})$ and $\frac{2}{|\mathbf{S}|+1}\sum_{F\in\mathbf{S}} I(F_i;F) < I(F_i;\mathbf{S})$ ($\frac{2}{|\mathbf{S}|+1}\sum_{F\in\mathbf{S}} I(F_i;F|C) \geq I(F_i;\mathbf{S}|C)$ and $\frac{2}{|\mathbf{S}|+1}\sum_{F\in\mathbf{S}} I(F_i;F|C) < I(F_i;\mathbf{S}|C)$) may all hold true under the different conditions.

Appendix A.2. Proof of Eq. (12)

With Eq. (2), the joint mutual information maximization (JMIM) criterion (Bennasar et al., 2015) can be written as

$$\begin{aligned} J_{JMIM}(F) &= \min_{F_\mathbf{S}\in\mathbf{S}} I(F,F_\mathbf{S};C) \\ &= I(F;C) + \min_{F_\mathbf{S}\in\mathbf{S}} [I(F_\mathbf{S};C) - I(F,F_\mathbf{S}) + I(F,F_\mathbf{S}|C)] \end{aligned} \tag{A.7}$$

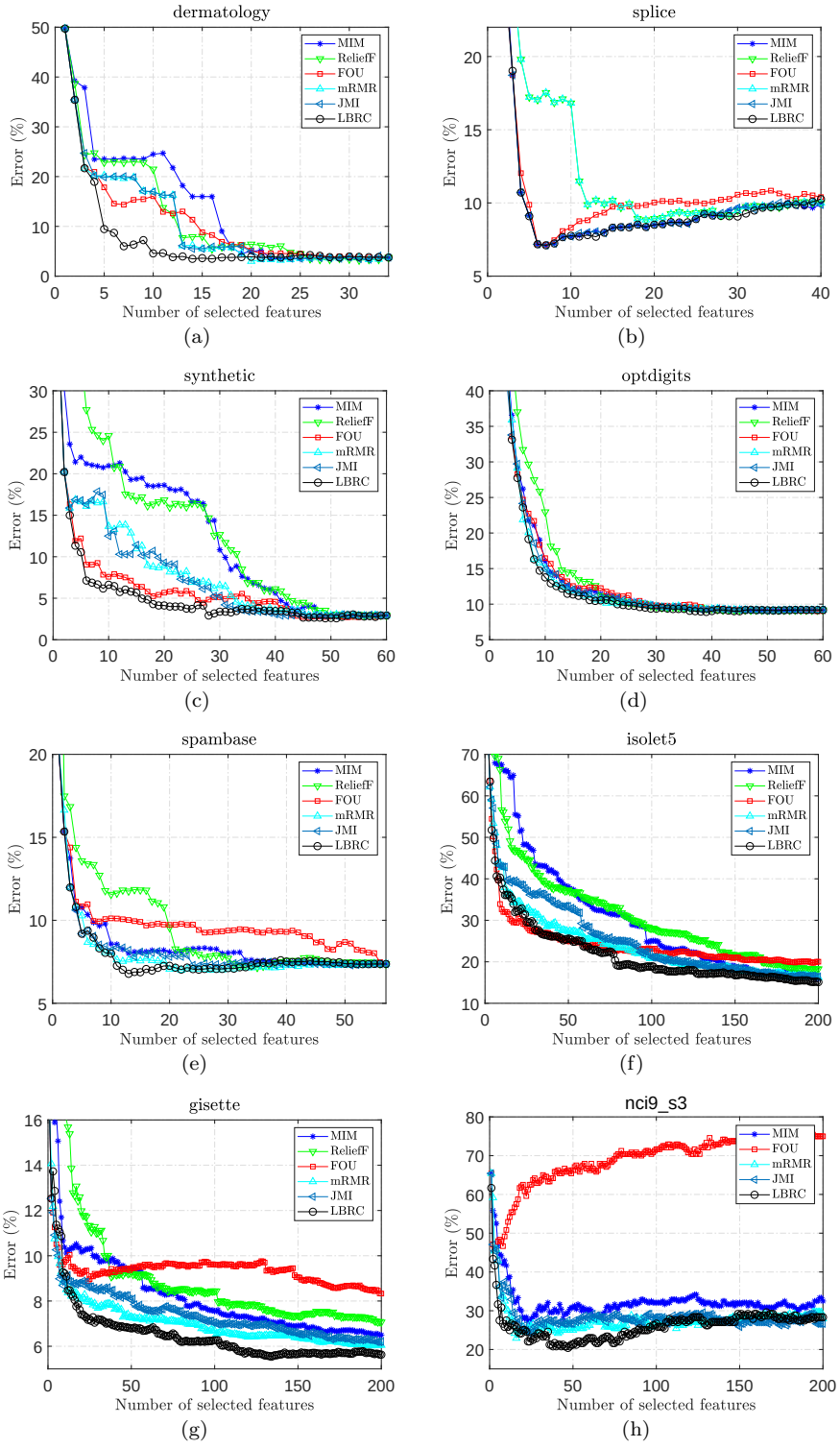

Figure 1: Classification error w.r.t. the number of selected features on UCI datasets. The result of ReliefF is inferior and out of scope and thus is omitted in 1(h).

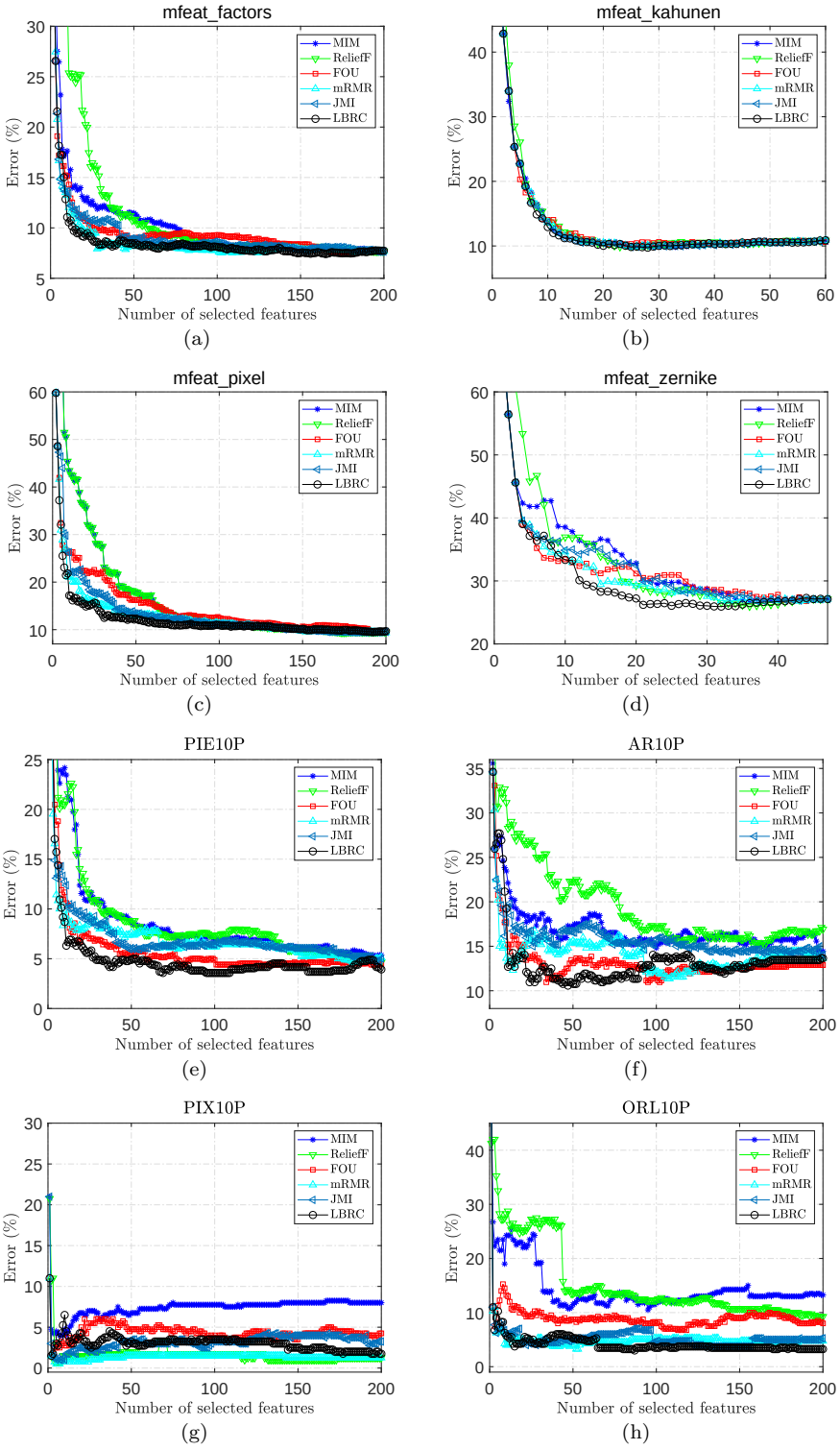

Figure 2: Classification error w.r.t. the number of selected features on image datasets.

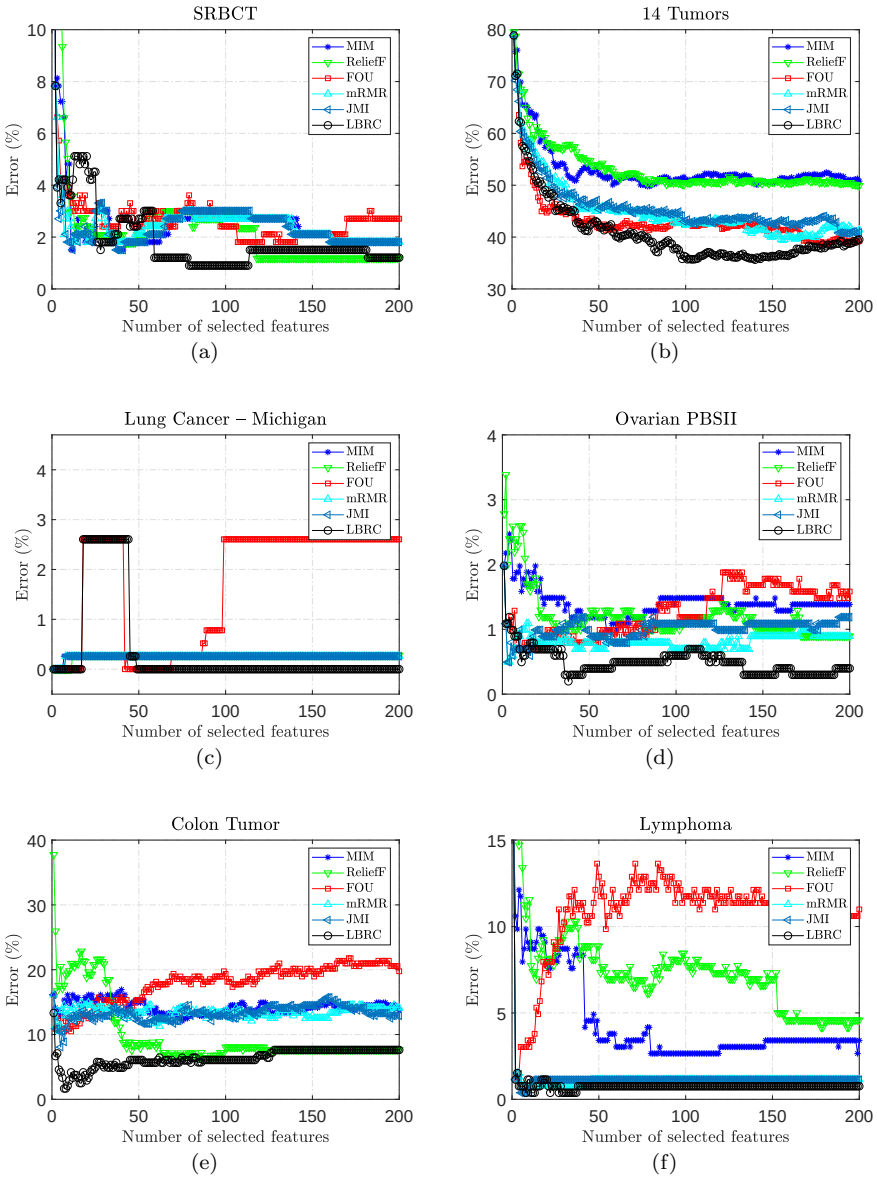

Figure 3: Classification error w.r.t. the number of selected features on microarray datasets.