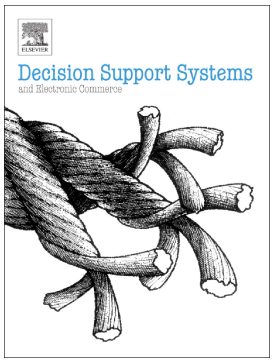

Feature assessment and ranking for classification with nonlinear sparse representation and approximate dependence analysis

Yishi Zhang, Qi Zhang, Zhijun Chen, Jennifer Shang, Haiying Wei

PII: S0167-9236(19)30080-6
DOI: https://doi.org/10.1016/j.dss.2019.05.004
Reference: DECSUP 13064

To appear in: *Decision Support Systems*

Received date: 10 December 2018
Revised date: 27 April 2019
Accepted date: 17 May 2019

Please cite this article as: Y. Zhang, Q. Zhang, Z. Chen, et al., Feature assessment and ranking for classification with nonlinear sparse representation and approximate dependence analysis, Decision Support Systems, https://doi.org/10.1016/j.dss.2019.05.004

Feature assessment and ranking for classification with nonlinear sparse representation and approximate dependence analysis

Yishi Zhang[a,d], Qi Zhang[b,*], Zhijun Chen[c,*], Jennifer Shang[d], Haiying Wei[a]

[a]*School of Management, Jinan University, Guangzhou 510632, China*
[b]*School of Economics and Management, China University of Geosciences (Wuhan), Wuhan 430074, China*
[c]*Intelligent Transport Systems Research Center, Wuhan University of Technology, Wuhan 430063, China*
[d]*Joseph M. Katz Graduate School of Business, University of Pittsburgh, Pittsburgh, PA 15260, USA*

Abstract

Feature selection has received significant attention in knowledge management and decision support systems in the past decades. In this study, kernel-based sparse representation and feature dependence analysis are integrated into a feature assessment and ranking framework. The proposed method utilizes the advantages of the kernel-based sparse representation technique and of the information theoretic metric to iteratively obtain the salient feature cluster. Then, a novel approximate dependence analysis is applied to further maintain complementarity while eliminating redundancy among the features selected by nonlinear orthogonal matching pursuit (NOMP). This can effectively prevent the significant bias caused by the pairwise correlation analysis for a large-scale feature set. To illustrate the effectiveness of the proposed method, classification experiments are conducted with three representative classifiers, on nine well-known datasets. The experimental results show the superiority of the proposed method compared with the representative information theoretic and model-based methods in classification for data-driven decision support systems.

Keywords: Feature selection; dimensionality reduction; classification; sparse representation; dependence analysis

1. Introduction

Selecting salient features that preserve or promote the performance of data mining and decision-making is a problem of growing significance because of the increasing size, high-dimensionality, and complexity of real-world datasets in numerous domains, e.g., financial decision-making and credit scoring (e.g. Serrano-Silva et al., 2018; Maldonado et al., 2017), image processing (e.g. Zhang et al., 2015a; Neoh et al., 2015; Zhang et al., 2016), and cancer diagnosis (Tan et al., 2018; Srisukkham et al., 2017). Taking the credit scoring in peer-to-peer lending as an example, automatically detecting which customers are possible to default on loan

*Corresponding authors.
Email addresses: yszhang@jnu.edu.cn (Yishi Zhang), qizhang_pitt@126.com (Qi Zhang), chenzj556@163.com (Zhijun Chen), shang@katz.pitt.edu (Jennifer Shang), tweihy@163.com (Haiying Wei)

repayment using machine learning methods has been a hot issue for years, as it can help financial experts to make profitable decisions while fulfilling regulatory requirements. However, tremendous financial features known as the *curse-of-dimensionality* will not only increase the computational cost but also impair the performance of the machine learning methods, owing to the inclusion of redundant and noisy information. Feature selection is thus applied to reduce the dimensionality of feature space and to boost the performance of the machine learning methods. Effective feature selection methods have been widely recognized for their capabilities in facilitating data acquisition, increasing learning efficiency, removing noise, and reducing overfitting.

Feature selection methods can be broadly classified into three distinct types: filter methods (e.g., Zhang et al., 2018a), wrapper methods (e.g., Kohavi and John, 1997), and embedded methods (e.g., Maldonado et al., 2017). Filter methods apply classifier-irrelevant metrics such as information theoretic metrics (Brown et al., 2012) and ℓ_p-norm ($0 < p \leq 2$) (Wang et al., 2010) to evaluate and select features; therefore, they generally incur relatively low computational cost compared with the following two types. Wrapper methods evaluate and select feature subsets based on specific classification accuracies. Thus, their performances are generally sensitive to the classifiers they use (Rodriguez-Lujan et al., 2010). In addition, the computational costs of wrapper methods are relatively high because each candidate of the feature subsets would be utilized to train the classifier. Embedded methods are classifiers where feature selection is integrated into the learning process (Maldonado et al., 2017). They are also classifier-specific, which limits their applications to other classifiers. Ordinarily, Filter is preferable to other types because of its superiority in numerous respects, e.g., remarkable generalization performance among different classifiers and high computational efficiency (Guyon and Elisseeff, 2003).

Filter methods developed in the early stage, such as mutual information maximization (MIM) (Lewis, 1992), evaluate and rank features in terms of only the relationship between the feature and the class (i.e., class-relevance). Such simplicity renders them highly efficient even in present day applications. However, they have a severe drawback: Features are evaluated individually, and the potential correlations among them that may severely influence the classification performance are not considered. Since it has been indicated that simply combining class-relevant features together cannot guarantee the reasonable performance of the learning method, a natural improvement is to additionally consider dependencies among features (Yu and Liu, 2004; Meyer and Bontempi, 2006; Brown et al., 2012). Because the combination of multiple correlations should be taken into account and the corresponding formulated problems are often nonconvex, it appears infeasible to obtain a globally optimal feature subset within polynomial time unless $\mathbf{P} = \mathbf{NP}$. Moreover, a reliable estimation of the joint distribution among features requires a large amount of samples, whereas in most of the real-world cases, samples are insufficient for even a medium-scale joint estimation. Therefore, a number of existing feature selection methods decompose the objective of feature selection into multiple sub-objectives, including maximizing class-relevance and minimizing feature inner-correlations (e.g., redundancy)

(Peng et al., 2005), and apply heuristic searching strategies and approximate dependence analysis (e.g., pairwise correlation analysis) for each sub-objective to finally obtain *satisfactory* solutions, i.e., to determine well-qualified feature subsets (e.g. Zhang et al., 2018a; Pandit et al., 2018) or feature sequences of which the top-ranked features are salient for data representation (e.g. Fleuret, 2004; Brown et al., 2012; Zhang et al., 2015b). Besides, explicit decomposing the objective of feature selection into multiple sub-objectives can improve the interpretability of the results for real-world applications, e.g., practitioners and empirical researchers can find potential collinearity by conducting redundancy analysis. However, most of the heuristic strategies employed in these methods seem to be intuitive and unsound. In addition, approximate strategies like pairwise correlation analysis may lead to significant bias in measuring feature dependencies. All these deficiencies result in an unstable performance of such methods.

Recently, sparse representation techniques have attracted increasing attention in numerous domains because they aim to obtain a small group of patterns to optimally recover the target, which can be formulated using a ℓ_0-norm objective or regularization term (Chen et al., 2010). In the context of feature selection, the aim of sparse representation is to determine a small number of features to preserve the classification accuracy (Efron et al., 2004; Wang et al., 2016; Peng et al., 2016). The most significant advantage of sparse representation-based feature selection is that, it provides a unifying and analytically solvable optimization framework for feature selection. Recent feature selection methods with sparse representation techniques utilize a variety of sparse models, such as the models with ℓ_1-norm (Liu and Zhang, 2016), ℓ_2-norm (Peng et al., 2016), and $\ell_{2,p}$-norm ($0 < p \leq 1$) (Tao et al., 2016), to select representative features. However, minimizing an ℓ_p-norm ($0 \leq p < 1$) regularized objective is proved to be strongly **NP**-hard (Chen et al., 2010). Although there are several relaxations, e.g., ℓ_1- or ℓ_2-norm as convex approximations, matrix computation incurs excessive execution time and space cost, and thus hinders the implementation of such methods on large-scale datasets. In addition, extant feature selection methods with sparse representation techniques do not explicitly handle feature inner-correlations, i.e., redundancy and complementarity (Chen et al., 2015), which is likely to severely influence the performance of the classification. This makes it similar to a *black box*, wherein it is infeasible to exactly determine whether or not sufficient efforts have been undertaken to handle feature inner-correlations.

Considering these, in this study, we propose a novel feature ranking method wherein sparse representation and information theoretic metrics are integrated to discriminate salient features, taking advantage of both sparse representation and feature inner-correlation analysis. More specifically, the proposed method not only utilizes the optimization framework of sparse representation, but also conducts dependence analysis to explicitly handle feature inner-correlations like redundancy and complementarity, in such a way as to obtain a salient and interpretable results for classification modeling. To our knowledge, this work is the first attempt to select features by combining sparse representation technique and information theoretic dependence analysis. The main contributions of this work that distinguishes it from extant literature are

threefold:

- A nonlinear sparse representation method is applied to identify representative feature clusters,
- conditional mutual information is used to identify the initial point for kernel orthogonal matching pursuit (OMP) in order to take into account the feature dependence, and
- a novel approximate dependence analysis strategy that can effectively prevent the significant bias caused by pairwise correlation analysis for large-scale feature set is proposed to eliminate redundancy and to keep complementarity.

The remainder of the paper is organized as follows: Section 2 introduces related work. Section 3 briefly describes feature inner-correlations in the context of information theory, the principle of sparse representation, and the overarching framework of the proposed method. Section 4 proposes a feature ranking approach based on kernel OMP. Then, section 5 proposes a novel approximate redundancy-complementarity analysis. The proposed method is presented in section 6. Section 7 presents the experimental results and discussions to evaluate the effectiveness of the proposed method in comparison with the representative feature selection methods on nine well-known datasets. Finally, section 8 summarizes the concluding remarks and indicates the future work.

2. Related Work

2.1. Feature selection with dependence analysis

The main objective of the present feature evaluation criteria considering feature dependencies is to identify a set of class-relevant and complementary features, wherein the redundancy among them is minimal. In general, feature dependencies comprise feature redundancy and feature complementarity (Zhang et al., 2014, 2015b), wherein redundancy has attracted significantly more attention than complementarity due to its detriment to classification performance. In order to identify class-relevance and redundancy, a number of novel feature-evaluation criteria have been developed (Zhang et al., 2018a; Fleuret, 2004; Peng et al., 2005; Meyer et al., 2008). For example, Zhang et al. (2018a) propose a novel improvement of the firefly algorithm (FA, see e.g. Zhang et al., 2018b) to effectively eliminate redundancy among the features. The near-optimal feature subsets obtained by their methods significantly outperform those obtained by the typical FA-based feature selection methods in classification and regression modeling. For information theoretic methods, the representative redundancy evaluation criterion is called minimum redundancy and maximum relevance (mRMR) (Peng et al., 2005). It applies mutual information (MI) to analyze the class-relevance for each feature and the correlation between any two features. Another example is available in Yu and Liu (2004) and Song et al. (2013), wherein fast correlation-based feature (FCBF) selection algorithms are developed

to separately handle class-relevance and redundancy using normalized mutual information. The conditional mutual information maximization (CMIM) criterion for feature selection is proposed by Fleuret (2004); here, conditional mutual information (CMI) is applied as the metric of feature dependencies. Because CMI and its equivalent variant joint mutual information (JMI) (Yang and Moody, 1999; Meyer et al., 2008) can jointly identify class-relevance and redundancy, they have been widely applied for feature assessment and selection in literature (Zhang et al., 2015b; Liang and Hu, 2017; Brown et al., 2012; Wang et al., 2017). For example, Meyer and Bontempi (2006) and Brown et al. (2012) both expand JMI into three dimensions, including class-relevance, redundancy, and interaction (in Meyer and Bontempi (2006), these three dimensions are multiplied by normalization coefficients to penalize inputs with large entropies), to explicitly handle redundancy and complementarity. Zhang et al. (2014) regard redundancy and complementarity as two poles of a comprehensive correlation which is called conditional redundancy, and then utilize CMI as the evaluation metric; Wang et al. (2017) expand CMI into three terms corresponding to class-relevance, redundancy, and complementarity, respectively, and assign different weights to the terms to heuristically select features with high class-relevance, high complementarity, and low redundancy. Despite of the relatively high efficiency, most of the above mentioned methods conduct dependence analysis by means of pairwise approximation (i.e., only measuring dependence between any two features), and lack reliable double-checking strategies trying to reduce the bias caused by pairwise approximation.

2.2. Feature selection with sparse representation techniques

Most extant feature selection works with sparse representation techniques are largely based on the formulations of least square regression (LSR) (Nie et al., 2010), linear discriminative analysis (LDA) (Tao et al., 2016), and support vector machines (SVM) (Thi et al., 2015; Peng et al., 2016). The most popular sparse regularization is ℓ_1-norm (namely Lasso) (Luo and Chen, 2014). Although feature selection based on ℓ_1-norm can select sparse features for its computational convenience, the results are often not sufficiently sparse. This implies that potential irrelevant and redundant features continue to be present in the selected feature subset. To date, there are two lines of research on sparse representation. One mainly focuses on improving the efficiency of sparse representation on high-dimensional data, where the typical method is called least angle regression (LARS) (Efron et al., 2004). The other aims to obtain sparser solutions. For example, a series of works (Foucart and Lai, 2009; Chen et al., 2010; Xu et al., 2012) extend ℓ_1-norm to ℓ_p-norm ($0 < p < 1$), and investigate its properties and possible applications. Xu et al. (2010) analyze $\ell_{1/2}$-norm regularization and indicate that it exhibits the highest performance among all the ℓ_p-norm regularizations with $p \in (0, 1)$. Then they use $\ell_{1/2}$-norm regularization for robust face recognition (Xu et al., 2012). However, the relaxation from ℓ_0- to ℓ_p-norm ($0 < p < 1$) cannot theoretically reduce the complexity of the original problem. Furthermore, Chen et al. (2014) show that ℓ_p-norm ($0 < p < 1$) minimization is also strongly **NP**-hard; this restricts its application in various fields, particularly where execution time plays

an essential role. From the perspective of a trade-off, the strong relaxations such as using ℓ_1- or ℓ_2-norms continue to be widely used in the fields where machine learning and data mining play important roles (Wang et al., 2016; Peng et al., 2016).

Regardless of the convex relaxations, matrix operations still require large execution time and space and thus are nearly intractable on large-scale datasets. For example, algorithms based on LSR and LDA with $\ell_{2,p}$-norm ($0 < p \leq 2$) regularization (Nie et al., 2010; Masaeli et al., 2010; Wang et al., 2010; Tao et al., 2016) all required the construction of a transformation matrix $\mathbf{D} \in \mathbb{R}^{m \times m}$ (m denotes the number of features in the feature space), resulting in large time and space overhead and thus hindering the use of such methods to address high-dimensionality.

3. Preliminaries and the Framework

3.1. Redundancy and complementarity from information theoretic perspective

As mentioned previously, the objective of feature selection can be factorized into interpretable sub-objectives that are necessary for dependence analysis using information theoretic metrics. In this section, we introduce certain necessary albeit fundamental information theoretic metrics that will be employed by the proposed method in the following context. Entropy, an essential quantitative description of information, is used to measure the extent of uncertainty for the distribution of a random variable X. It has the following formulation:

$$H(X) = -\int_X p(x) \log p(x) \mathrm{d}x, \tag{1}$$

where x is the possible value assignment of X, $p(\cdot)$ is the probability density function (for convenience, we hereafter use the notation log to denote logarithm to the base two). According to information theory, conditional entropy could be used to quantify the uncertainty of a variable conditioned on another. The conditional entropy of X given Y is defined as

$$H(X|Y) = -\iint_{XY} p(xy) \log p(x|y) \mathrm{d}x\mathrm{d}y, \tag{2}$$

Then, we can formulate the mutual information (MI) between two random variables X and Y:

$$\begin{aligned} I(X;Y) &= \iint_{XY} p(xy) \log \frac{p(xy)}{p(x)p(y)} \mathrm{d}x\mathrm{d}y \\ &= H(X) - H(X|Y), \end{aligned} \tag{3}$$

where x and y are the possible value assignments of X and Y, respectively. MI can be considered as the amount of information shared by two variables. In the context of feature selection, MI is one of the most widely used metrics for measuring the correlation intensity of two features. Note that MI is symmetrical, i.e., $I(X;Y) = I(Y;X)$. $I(X;Y) = 0$ implies that X and Y are statistically independent.

Estimating MI between two random variables given the third can be achieved by CMI, which is widely used in information theoretic learning methods. CMI is defined as

$$\begin{aligned} I(X;Y|Z) &= \iiint_{XYZ} p(xyz)\log\frac{p(xy|z)}{p(x|z)p(y|z)}\mathrm{d}x\mathrm{d}y\mathrm{d}z \\ &= H(X|Z) - H(X|Y,Z). \end{aligned} \tag{4}$$

$I(X;Y|Z)$ can be interpreted as the information shared by X and Y given the value of a third variable Z. CMI is also symmetrical for X and Y. A remarkable property of CMI in feature evaluation is that it can effectively distinguish *useful* features from *useless* ones. A large value of $I(X;Y|Z)$ implies a strong relevance between X and Y when the distribution of Z is known, whereas a small $I(X;Y|Z)$ implies negligible relevance when the distribution of Z is known, regardless of the original correlation between X and Y. For feature selection, the latter advantage prevents unnecessary efforts for further identifying whether X and Y are originally irrelevant or redundant when Z is given.

According to Brown et al. (2012) and Chen et al. (2015), CMI can simultaneously capture three essential feature correlations: class-relevance, redundancy, and complementarity among features. This can be demonstrated by expanding CMI as

$$I(F_1;C|F_2) = I(F_1;C) - I(F_1;F_2) + I(F_1;F_2|C). \tag{5}$$

Herein, the first term on the right-hand side of Eq. (5) captures the class-relevance of F_1, the second captures the redundancy between F_1 and F_2, and the final term captures the complementary correlation between F_1 and F_2. This renders CMI a highly effective metric that can simultaneously give attention to the three feature correlations that play crucial roles in feature selection.

3.2. Sparse representation and orthogonal matching pursuit

In this section, we briefly introduce the fundamental formula of sparse representation and an effective solver named orthogonal matching pursuit (OMP) that will be applied in the proposed method. Recently, sparse representation has been playing increasingly important roles in feature selection and sparse reconstruction (Foucart and Lai, 2009; Chen et al., 2010, 2014). In general, sparse representation can be formalized as the determination of an optimal solution of a minimization programming wherein the objective function is the sum of a data-fitting term in ℓ_2-norm and a regularization term in ℓ_0-norm. As ℓ_0-norm minimization is strongly **NP**-hard, its convex relaxation using ℓ_p-norm ($p \geq 1$) is often applied in literature (Chen et al., 2014). With a slight abuse of notation, let $\mathbf{F} = (F_1, ..., F_k)$ (where k is the number of features) and the i-th feature $F_i = (f_{i,i}, ..., f_{n,i})^T$ (where $f_{j,i}$ is the value of F_i in the j-th sample and n is the number of samples). Then, a linear sparse representation of the class C can be ideally described by all the features as

$$C = \beta_1 \cdot \begin{pmatrix} f_{1,1} \\ ... \\ f_{n,1} \end{pmatrix} + \beta_2 \cdot \begin{pmatrix} f_{1,2} \\ ... \\ f_{n,2} \end{pmatrix} + ... + \beta_k \cdot \begin{pmatrix} f_{1,k} \\ ... \\ f_{n,k} \end{pmatrix}. \tag{6}$$

Herein, $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_k)^T \in \mathbb{R}^k$ is a coefficient vector where most of the entries are expected to be zero. This can be obtained by solving the following minimization programming

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^k} \ ||\boldsymbol{\beta}||_0, \quad \text{s.t. } \mathbf{F} \cdot \boldsymbol{\beta} = C, \tag{7}$$

where $|| \cdot ||_0$ denotes the number of nonzero entries in the norm. However, model (7) is strongly **NP**-hard, rendering it infeasible to be exactly solved even within pseudo-polynomial time unless $\mathbf{P} = \mathbf{NP}$. Recent development in compressed sensing indicates that if the exact solution is sufficiently sparse, it can be obtained by solving the following relaxed programming with $p = 1$ (Candes and Tao, 2005):

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^k} \ ||\boldsymbol{\beta}||_p = \left(\sum_{i=1}^{k} |\beta_i|^p \right)^{1/p}, \quad \text{s.t. } \mathbf{F} \cdot \boldsymbol{\beta} = C. \tag{8}$$

The objective of model (8) is convex when $p = 1$, and thus, optimal $\boldsymbol{\beta}^*$ can be achieved within polynomial time. Because real-world datasets ordinarily have redundancy among the samples (i.e., sample matrices are not fully ranked), the equivalent relationship in the constraint of model (8) cannot strictly hold. A dense noise ψ is hence introduced and the equation becomes $C = \mathbf{F} \cdot \boldsymbol{\beta} + \psi$. Now, the problem is transformed to

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^k} \ ||\boldsymbol{\beta}||_p = \left(\sum_{i=1}^{k} |\beta_i|^p \right)^{1/p}, \quad \text{s.t. } ||\mathbf{F} \cdot \boldsymbol{\beta} - C||_2 \leq \epsilon \tag{9}$$

where ϵ is the upper bound of $||\psi||_2$. The Lagrangian form of Model (9) is represented as

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^k} \ ||\mathbf{F} \cdot \boldsymbol{\beta} - C||_2^2 + \lambda \cdot ||\boldsymbol{\beta}||_p^p, \tag{10}$$

where λ is the regularized coefficient. Model (10) is the style of the popular Lasso when $p = 1$. As a trade-off between runtime and accuracy, a greedy algorithm called orthogonal matching pursuit (OMP) that performs reasonably in sparse representation in various fields (Cai and Wang, 2011) can be utilized to solve (10). Before we describe the process of OMP, we denote Ω as the index vector comprising the indices of the current selected features from $\mathbf{F}$ and $\mathbf{F}_\Omega$ as the feature subset (the submatrix of $\mathbf{F}$) where the entries (columns) are the selected features from $\mathbf{F}$ indexed by Ω, with a marginal abuse of notation. We use the superscript to denote the iteration step and the subscript to denote the index of the entry (column) in the set (matrix).

OMP is a stepwise forward selection algorithm and is simple to implement. However, model (10) can only capture linear correlation in real-world tasks. Because the effectiveness of the linear sparsity model largely depends on the structure of the data, it may not enable sufficiently accurate dimension reduction to be reliable if the data is not sufficiently linearly-separable. Kernel-based algorithms (Mo et al., 2014) implicitly exploit the nonlinear structure of the data. Kernel methods can be applied to transform the data into a higher dimensional space via a transformation $\phi(\cdot)$ such that the resulting transformed data becomes

Input: $\mathbf{F}$ /* feature set */, C /* class */

Output: $\mathbf{S}$ /* selected feature subset */, Ω /* Indices of selected features in $\mathbf{F}$ */

Initialize $\mathbf{S} \leftarrow \varnothing$, $\Omega^0 \leftarrow \varnothing$, residual $\mathbf{r}^0 = C$, $i \leftarrow 1$ /* iteration counter */

repeat

  Greedy search: Find the feature $\widetilde{F}_j$ satisfying $\max_{F \in \mathbf{F}} \{F^T \cdot \mathbf{r}^{i-1}\}$

  $\Omega^i = \Omega^{i-1} \cup \{j\}$

  Projection: $\mathbf{P}_i \leftarrow \mathbf{F}_{\Omega^i} \cdot \left(\mathbf{F}_{\Omega^i}^T \cdot \mathbf{F}_{\Omega^i}\right)^{-1} \cdot \mathbf{F}_{\Omega^i}^T$ /* $\mathbf{P}_i$ denotes the projection onto the linear space spanned by the elements of $\mathbf{F}_{\Omega^i}$ */

  Update current residual: $\mathbf{r}^i \leftarrow (\mathbf{I} - \mathbf{P}_i) \cdot C$ /* $\mathbf{I}$ denotes the identity matrix. */

  $i \leftarrow i + 1$

until the stopping criterion is satisfied, i.e., $\frac{||\mathbf{r}^i||_2}{||\mathbf{r}^{i-1}||_2} \geq 0.95$;

$\mathbf{S} \leftarrow \mathbf{F}_{\Omega^i}$

return $\mathbf{S}$

Algorithm 1: Orthogonal matching pursuit

more separable and comply with the linear sparsity model. In this work, kernel operators are embedded in model (10) and thus render the model more robust in real-world feature selection tasks.

Given the transformed vectors $\mathbf{x} \mapsto \phi(\mathbf{x})$ and $\mathbf{y} \mapsto \phi(\mathbf{y})$, the kernel function $\kappa : \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}$, can be defined as the inner product of the transformed vectors: $\kappa(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^T \cdot \phi(\mathbf{y})$. The linear sparsity model (10) can be transformed as

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^k} \ ||\mathbf{F}_\phi \cdot \boldsymbol{\beta} - \phi(C)||_2^2 + \lambda \cdot ||\boldsymbol{\beta}||_p^p, \tag{11}$$

where $\mathbf{F}_\phi = (\phi(F_1), ..., \phi(F_k))$. Gaussian radial basis function (RBF) with the form

$$\kappa(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{||\mathbf{x} - \mathbf{y}||_2}{2\sigma^2}\right) \tag{12}$$

is mostly applied, and it performs effectively in real-world applications (Mo et al., 2014). In the following section, Gaussian RBF will be applied as the kernel function and Algorithm 1 will be modified to solve model (11) for the feature clustering task.

3.3. The proposed framework

Effective data-driven decision support systems require the raw data from multiple sources to be efficiently gathered and pre-processed for redundancy and noise elimination to facilitate classification modeling and enhance the performance for undertaking data-driven decision-making processes. Fig. 1 provides the diagram of the decision support process that integrates the framework of our proposed feature assessment and ranking method.

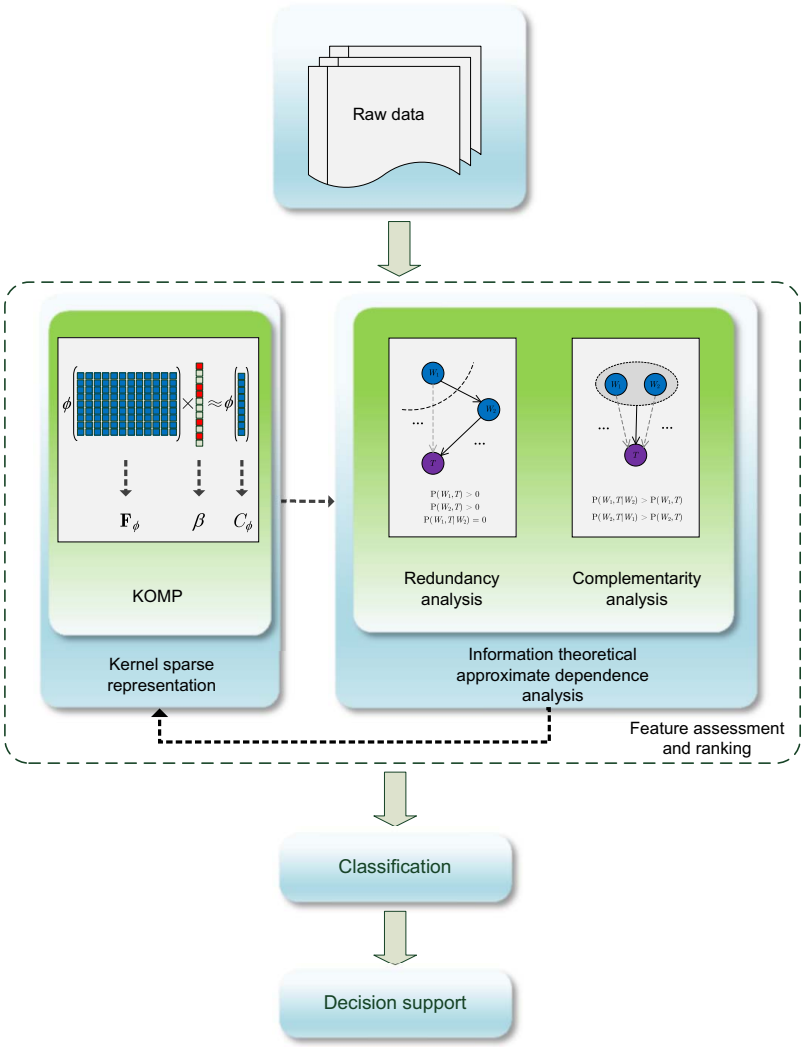

Figure 1: The process diagram for data-driven decision support processes.

The framework of the proposed method shown in Fig. 1 consists of two main submodules. The first submodule (shown in left-hand side of Fig. 1) focuses on feature clustering. That is, it utilizes sparse representation techniques to categorize features into *representative* and *non-representative* groups according to a certain task. The representative features are candidates for final-selected features with high importance weights, while the non-representative features are weighted relatively lighter but still possible to be identified as representative in the next iterations. The second submodule (shown in right-hand side of Fig. 1) conducts dependence analysis for the representative features obtained from the first submodule for redundancy elimination and complementarity identification using information theoretic metrics. The features identified as redundant in this stage will be finally eliminated, because they may impair the discriminative power of other representative features with higher importance weights. Those identified as relevant or complementary

will be assessed as salient features and finally top-ranked among the selected features.

Next, we introduce in detail the sparse representation technique and the information theoretic dependence analysis applied in the proposed method, and the collaborative feature selection process of the two submodules.

4. Feature Clustering by Sparse Representation

Sparse representation techniques can effectively select a small part of the features to represent the class. That is, features are separated into two groups: one group contains the selected relevant features for class representation, and the other contains the rest. Thus, sparse representation techniques are preeminent choices as clustering methods for class-relevance analysis. This task can be effectively formulated using model (11). Regarding the nonlinearity of the feature space, the kernel operator is a better choice to be applied in our method. Accordingly, OMP is required to be modified to nonlinear orthogonal matching pursuit (NOMP) to effectively capture the nonlinear correlations among the features. Recall that the inner product computation is conducted three times in Algorithm 1, namely, greedy search (line 3 in Algorithm 1), projection (line 5), and residual updating (line 6). The kernel operator thus replaces the inner product and is integrated into each of them as follows. For convenience, certain notations in the Matlab syntax will be used in the following context. Let $\mathbf{\Gamma}_\phi \in \mathbb{R}^k$ be the kernel vector of which the j-th entry $\mathbf{\Gamma}_\phi(j) = \kappa(F_j, C)$, and $\mathbf{\Lambda}_\phi \in \mathbb{R}^{k\times k}$ be the kernel matrix of which the (j,l)-th entry $\mathbf{\Lambda}_\phi(j,l) = \kappa(F_j, F_l)$. $\mathbf{\Gamma}_\phi(\Omega)$ denotes the vector of which the entries are the corresponding ones in $\mathbf{\Gamma}_\phi$ indexed by the entries of Ω. $\mathbf{\Lambda}_\phi(\Omega,\Omega)$ denotes the matrix of which the columns and rows are the corresponding ones in $\mathbf{\Lambda}_\phi$ indexed by the entries of Ω, respectively. Then, the projection matrix $\mathbf{P}$ can be formulated as

$$\begin{aligned}\mathbf{P} &= [\phi(F_{\Omega_1}), ..., \phi(F_{\Omega_{|\Omega|}})] \cdot \left([\phi(F_{\Omega_1}), ..., \phi(F_{\Omega_{|\Omega|}})]^T \cdot [\phi(F_{\Omega_1}), ..., \phi(F_{\Omega_{|\Omega|}})]\right)^{-1} \cdot [\phi(F_{\Omega_1}), ..., \phi(F_{\Omega_{|\Omega|}})]^T \\ &= [\phi(F_{\Omega_1}), ..., \phi(F_{\Omega_{|\Omega|}})] \cdot \mathbf{\Lambda}_\phi^{-1}(\Omega,\Omega) \cdot [\phi(F_{\Omega_1}), ..., \phi(F_{\Omega_{|\Omega|}})]^T \end{aligned} \quad (13)$$

The residual updating process can be formulated as

$$\begin{aligned}\phi(\mathbf{r}) &= (\mathbf{I} - \mathbf{P}) \cdot \phi(C) \\ &= \phi(C) - [\phi(F_{\Omega_1}), ..., \phi(F_{\Omega_{|\Omega|}})] \cdot \mathbf{\Lambda}_\phi^{-1}(\Omega,\Omega) \cdot [\phi(F_{\Omega_1}), ..., \phi(F_{\Omega_{|\Omega|}})]^T \cdot \phi(C). \end{aligned} \quad (14)$$

Since $\mathbf{P}^2 = \mathbf{P}$, we can get

$$\begin{aligned}
\kappa(\mathbf{r},\mathbf{r}) &= \phi(\mathbf{r})^T \cdot \phi(\mathbf{r}) \\
&= \kappa(C,C) + \phi(C)^T \cdot \left(\mathbf{P}^2 - 2\mathbf{P}\right) \cdot \phi(C) \\
&= \kappa(C,C) - \phi(C)^T \cdot [\phi(F_{\Omega_1}), ..., \phi(F_{\Omega_{|\Omega|}})] \cdot \mathbf{\Lambda}_\phi^{-1}(\Omega,\Omega) \cdot [\phi(F_{\Omega_1}), ..., \phi(F_{\Omega_{|\Omega|}})]^T \cdot \phi(C) \\
&= \kappa(C,C) - \underbrace{[\kappa(C,F_{\Omega_1}), ..., \kappa(C,F_{\Omega_{|\Omega|}})]}_{\mathbf{\Gamma}_\phi^T(\Omega)} \cdot \mathbf{\Lambda}_\phi^{-1}(\Omega,\Omega) \cdot \underbrace{[\kappa(F_{\Omega_1},C), ..., \kappa(F_{\Omega_{|\Omega|}},C)]^T}_{\mathbf{\Gamma}_\phi(\Omega)} \\
&= \kappa(C,C) - \mathbf{\Gamma}_\phi^T(\Omega) \cdot \mathbf{\Lambda}_\phi^{-1}(\Omega,\Omega) \cdot \mathbf{\Gamma}_\phi(\Omega). \qquad (15)
\end{aligned}$$

For the greedy search process, suppose it is in the i-th step; then, we obtain

$$\begin{aligned}
\kappa(F_j^T, \mathbf{r}^i) &= \phi(F_j)^T \cdot \phi(\mathbf{r}^i) \\
&= \phi(F_j)^T \cdot \phi(C) - \phi(F_j)^T \cdot [\phi(F_{\Omega_1^i}), ..., \phi(F_{\Omega^i_{|\Omega^i|}})] \cdot \mathbf{\Lambda}_\phi^{-1}(\Omega^i,\Omega^i) \cdot [\phi(F_{\Omega_1^i}), ..., \phi(F_{\Omega^i_{|\Omega^i|}})]^T \cdot \phi(C) \\
&= \underbrace{\kappa(F_j,C)}_{\mathbf{\Gamma}_\phi(j)} - \underbrace{[\kappa(F_j,F_{\Omega_1^i}), ..., \kappa(F_j,F_{\Omega^i_{|\Omega^i|}})]}_{\mathbf{\Lambda}_\phi(j,\Omega^i)} \cdot \mathbf{\Lambda}_\phi^{-1}(\Omega^i,\Omega^i) \cdot \underbrace{[\kappa(F_{\Omega_1^i},C), ..., \kappa(F_{\Omega^i_{|\Omega^i|}},C)]^T}_{\mathbf{\Gamma}_\phi(\Omega^i)} \\
&= \mathbf{\Gamma}_\phi(j) - \mathbf{\Lambda}_\phi(j,\Omega^i) \cdot \mathbf{\Lambda}_\phi^{-1}(\Omega^i,\Omega^i) \cdot \mathbf{\Gamma}_\phi(\Omega^i), \qquad (16)
\end{aligned}$$

Equation (16) illustrates a straightforward implementation of kernel-based OMP: It renders the greedy search executable without the knowledge of the values of $\phi(F_j)$ and $\phi(\mathbf{r}^i)$. It should be also noted that the distance captured by $||F-C||_2^2$ is likely to be excessively large for $F \in \mathbf{F}$ because of dimensional inconsistency between the features and the class. This may result in accuracy loss or even invalid computational results when it is implemented using computing platforms such as Matlab. To achieve this, we apply $||\mathbf{F}||_F$ as the modified factor and use $\frac{||F-C||_2^2}{||\mathbf{F}||_F}$ rather than $||F-C||_2^2$ in the proposed method. We give the pseudo code of NOMP in Algorithm 2.

The stop rule of NOMP applied in Algorithm 2 depends on the noise structure of the data (Cai and Wang, 2011). In the noiseless case, the rule should be $\mathbf{r} = 0$. In this work, the convergence of NOMP is measured as the change in the residual $\mathbf{r}$, and we heuristically set the stop rule as $\frac{||\phi(\mathbf{r}^i)||_2}{||\phi(\mathbf{r}^{i-1})||_2} \geq 0.95$. In addition, it should be noted that $\mathbf{\Lambda}_\phi(\Omega^i,\Omega^i)$ cannot be guaranteed to be nonsingular in many cases. Since a valid inversion can be conducted only on nonsingular matrices, a regularization term is introduced by adding a constant value to the diagonal elements of $\mathbf{\Lambda}_\phi(\Omega^i,\Omega^i)$, i.e., $\mathbf{\Lambda}_\phi(\Omega^i,\Omega^i) + \lambda \cdot \mathbf{I} \quad (\lambda > 0)$. λ is generally a small number and is set as 0.0005 in the proposed method.

5. Approximate Dependence Analysis

As mentioned previously, feature selection methods with sparse representation do not explicitly handle redundancy and complementarity. It seems hard to know exactly whether or not those methods undertake

Input: $\mathbf{F}$ /* feature set */, C /* class */

Output: $\mathbf{S}$ /*selected feature subset*/, Ω /* Indices of selected features in $\mathbf{F}$ */

Initialize: Calculate $\mathbf{\Gamma}_\phi$ and $\mathbf{\Lambda}_\phi$, $\mathbf{r}^0 \leftarrow C$, $\mathbf{S} \leftarrow \varnothing$, $\Omega^0 \leftarrow \varnothing$, $i \leftarrow 1$

Calculate $||\phi(\mathbf{r}^0)||_2 = \sqrt{\kappa(C, C)}$

$\Omega^1 \leftarrow \Omega^0 \cup \{j | \arg\max_j \mathbf{\Gamma}_\phi(j)\}$

repeat

  Find the feature $\widetilde{F}_j$ satisfying the maximization problem:

  $\max_{F \in \mathbf{F}} \ \{\mathbf{\Gamma}_\phi(j) - \mathbf{\Lambda}_\phi(j, \Omega^i) \cdot \left(\mathbf{\Lambda}_\phi(\Omega^i, \Omega^i) + \lambda \cdot \mathbf{I}\right)^{-1} \cdot \mathbf{\Gamma}_\phi^T(\Omega^i)\}$

  $\Omega^{i+1} = \Omega^i \cup \{j\}$

  Calculate $||\phi(\mathbf{r}^i)||_2 = \sqrt{\kappa(\mathbf{r}^i, \mathbf{r}^i)}$ in terms of Eq.(15)

  $i \leftarrow i + 1$

until the stopping criterion is satisfied, i.e., $\frac{||\phi(\mathbf{r}^i)||_2}{||\phi(\mathbf{r}^{i-1})||_2} \geq 0.95$;

$\mathbf{S} \leftarrow \mathbf{F}_{\Omega^{i+1}}$

return $\mathbf{S}$

Algorithm 2: Nonlinear orthogonal matching pursuit

sufficient efforts to handle feature inner-correlations. On the contrary, information theoretic feature evaluation strategies generally achieve remarkable performances in redundancy and complementarity analysis and thus are included in this work to cover the deficiency of sparse representation techniques with respect to these aspects.

Because highly correlated features are likely to exhibit similar discriminative power, taking into account only the relevance between the features and the class is likely to result in two *neglects*, which possibly impair the quality of the selected features for classification (Chen et al., 2015): One is the neglect of redundancy among the selected features, caused by the assumption of individual independence of features (e.g. MIM). The other is the neglect of group capacity of features, caused by only measuring pairwise redundancy among features (e.g. mRMR). For example, features with low individual capacity are identified as either irrelevant or redundant by measuring their pairwise correlation. However, it is also likely that some of those features contribute largely to the discriminative power of the whole feature subset if they are selected. Thus, feature complementarity becomes a conspicuous correlation among features that should be applied to partially prevent the bias caused by pairwise redundancy analysis (Sun et al., 2013; Chen et al., 2015). However, such a complementarity analysis in literature still focuses on pairwise correlation of features, whereas k-wise ($k \geq 3$) complementary correlation among features are omitted; this is likely to also result in bias in feature evaluation, particularly when the size of the selected feature subset is large.

In this section, we focus on dependence analysis for the features generated within each execution of

NOMP. The highlight of this strategy is that it can effectively prevent bias caused by the pairwise correlation analysis, because the feature subset (cluster) selected from each iteration is almost constrained in a small scale (owing to the effective sparse representation of the features). We start our analysis from the perspective of redundancy–complementarity dimension because redundancy and complementarity can be measured comprehensively as two dimensions of feature dependence (Chen et al., 2015). The detailed procedure of the redundancy–complementarity analysis in the proposed method is depicted in Fig. 2. Specifically, the procedure first identifies the dependence of the features from the perspective of the redundancy–complementarity dimension: If complementarity arises, the features are selected; otherwise, the dominance of the features (i.e., which feature contributes more to the discriminative capability) is further analyzed to determine which one is redundant. In addition, to partially overcome the estimation bias arising from the finite samples, a significance rule is finally introduced to eliminate redundancy.

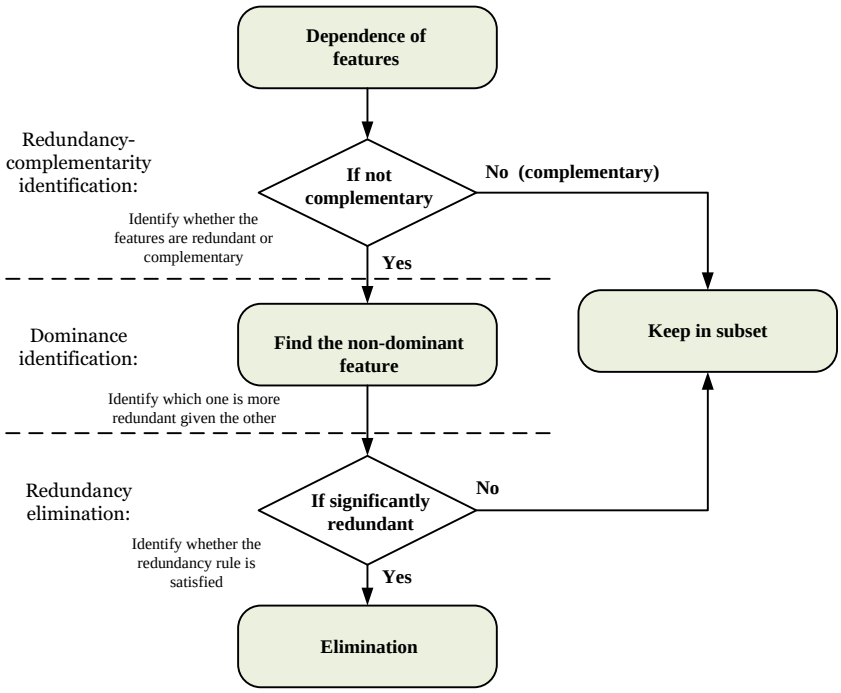

Figure 2: Process diagram of dependence analysis in the proposed method.

5.1. Redundancy-complementarity identification

To illustrate the dependence identification process in detail, we first propose certain equivalent transformations of MI and CMI in the following theorem.

Theorem 1. *For $\forall F_i, F_j \in \mathbf{F}$ $(F_i \neq F_j)$ and the class C, we have*

$$I(F_iF_j;C) \geq I(F_i;C) + I(F_j;C) \tag{17}$$

$$\Longleftrightarrow I(F_i;C) \leq I(F_i;C|F_j) \tag{18}$$

$$\Longleftrightarrow I(F_j;C) \leq I(F_j;C|F_i). \tag{19}$$

Proof. Without loss of generality, we only need to prove

$$I(F_iF_j;C) \geq I(F_i;C) + I(F_j;C) \Longleftrightarrow I(F_i;C) \leq I(F_i;C|F_j).$$

Since

$$\begin{aligned} &I(F_i;F_j) - I(F_i;F_j|C) \\ &= \iint_{F_iF_j} p(f_if_j)\log\frac{p(f_if_j)}{p(f_i)p(f_j)}\mathrm{d}f_i\mathrm{d}f_j \qquad (20)\\ &\quad - \iiint_{F_iF_jC} p(f_if_jc)\log\frac{p(f_if_j|c)}{p(f_i|c)p(f_j|c)}\mathrm{d}f_i\mathrm{d}f_j\mathrm{d}c \qquad (21)\end{aligned}$$

and with the fact

$$p(f_if_j) = \int_C p(f_if_jc)\mathrm{d}c,$$

Eq. (20) can be rewritten as

$$\begin{aligned} &I(F_i;F_j) \\ &= \iiint_{F_iF_jC} p(f_if_jc)\log\frac{p(f_if_j)}{p(f_i)p(f_j)}\mathrm{d}f_i\mathrm{d}f_j\mathrm{d}c \qquad (22)\end{aligned}$$

With Eqs. (21) and (22), we have

$$\begin{aligned} &I(F_i;F_j) - I(F_i;F_j|C) \\ &= \iiint_{F_iF_jC} p(f_if_jc)\log\left(\frac{p(f_if_j)}{p(f_i)p(f_j)}\cdot\frac{p(f_i|c)p(f_j|c)}{p(f_if_j|c)}\right)\mathrm{d}f_i\mathrm{d}f_j\mathrm{d}c \\ &= \iiint_{F_iF_jC} p(f_if_jc)\log\frac{p(f_if_j)p(f_ic)p(f_jc)}{p(f_i)p(f_j)p(c)p(f_if_jc)}\mathrm{d}f_i\mathrm{d}f_j\mathrm{d}c \\ &= \iiint_{F_iF_jC} p(f_if_jc)\log\left(\frac{p(f_ic)}{p(f_i)p(c)}\cdot\frac{p(f_if_j)p(f_jc)}{p(f_if_jc)p(f_j)}\right)\mathrm{d}f_i\mathrm{d}f_j\mathrm{d}c \\ &= \iint_{F_iC} p(f_ic)\log\frac{p(f_ic)}{p(f_i)p(c)}\mathrm{d}f_i\mathrm{d}c \\ &\quad - \iiint_{F_iF_jC} p(f_if_jc)\log\frac{p(f_ic|f_j)}{p(f_i|f_j)p(c|f_j)}\mathrm{d}f_i\mathrm{d}f_j\mathrm{d}c \\ &= I(F_i;C) - I(F_i;C|F_j),\end{aligned}$$

i.e.

$$I(F_i;F_j) - I(F_i;F_j|C) = I(F_i;C) - I(F_i;C|F_j), \tag{23}$$

we have

$$I(F_i; C|F_j) = I(F_i; C) - I(F_i; F_j) + I(F_i; F_j|C). \tag{24}$$

According to the chain rule of CMI (Huang and Chow, 2005), we have

$$I(F_iF_j; C) = I(F_j; C) + I(F_i; C|F_j). \tag{25}$$

With Eqs. (24) and (25), we obtain

$$I(F_iF_j; C) = I(F_i; C) + I(F_j; C) - I(F_i; F_j) + I(F_i; F_j|C). \tag{26}$$

Therefore,

$$\begin{aligned} & I(F_iF_j; C) \geq I(F_i; C) + I(F_j; C) \\ \Longleftrightarrow\ & I(F_i; F_j) \leq I(F_i; F_j|C). \end{aligned} \tag{27}$$

With Eq. (23), we finally obtain

$$\begin{aligned} & I(F_iF_j; C) \geq I(F_i; C) + I(F_j; C) \\ \Longleftrightarrow\ & I(F_i; C) \leq I(F_i; C|F_j). \end{aligned}$$

The proof is complete. □

In Theorem 1, $I(F_iF_j; C)$ captures the correlation between the feature group $\{F_i, F_j\}$ and the class C. From the perspective of the redundancy–complementarity dimension, $I(F_iF_j; C) > I(F_i; C) + I(F_j; C)$ reveals that the joint discriminative capability of $\{F_i, F_j\}$ is enhanced when they are drawn together; this implies complementarity between F_i and F_j. On the contrary, $I(F_iF_j; C) < I(F_i; C) + I(F_j; C)$ reveals that the joint discriminative capability of $\{F_i, F_j\}$ is impaired when they are drawn together; this implies redundancy between F_i and F_j. Theorem 1 provides an alternative perspective to identify complementarity and redundancy: If the appearance of F_i increases the relevance between F_j and the class, F_i and F_j are complementary to each other. Otherwise, redundancy is present and needs to be further analyzed. Thus, we can only compare $I(F_i; C)$ and $I(F_i; C|F_j)$ (or $I(F_j; C)$ and $I(F_j; C|F_i)$) for redundancy–complementarity analysis rather than compare the joint mutual information and the summation of the individual mutual information.

5.2. Dominance identification

This step focuses on identifying the dominant feature from the pair of features. That is, it would label one feature as *redundant*, which may finally be eliminated. According to Theorem 1, redundancy implies $I(F_i; C) > I(F_i; C|F_j)$ and $I(F_j; C) > I(F_j; C|F_i)$. Thus, the difference between $I(F_i; C)$ and $I(F_i; C|F_j)$

(and that between $I(F_j;C)$ and $I(F_j;C|F_i)$) can be applied to measure the magnitude of redundancy. We define the redundancy rate of F_i given F_j as

$$\Delta(F_i|F_j) = \frac{I(F_i;C) - I(F_i;C|F_j)}{I(F_i;C)}, \tag{28}$$

then, the dominant feature can be identified by comparing $\Delta(F_i|F_j)$ and $\Delta(F_j|F_i)$ since $\Delta(\cdot)$ could measure the extent to which the discriminative capability of a feature can be substituted by that of another. To conduct this more efficiently, we propose the following theorem:

Theorem 2. *For $\forall F_i, F_j \in \mathbf{F}$ $(F_i \neq F_j)$ and the class C, if $I(F_iF_j;C) \geq I(F_i;C) + I(F_j;C)$, we obtain*

$$\frac{\Delta(F_i|F_j)}{\Delta(F_j|F_i)} = \frac{I(F_j;C)}{I(F_i;C)}.$$

The proof of Theorem 2 is omitted as it is straightforward with Eq. (23). Theorem 2 reveals that the redundancy rate of a feature is proportional to its class-relevance. It provides a surprisingly efficient means to identify dominance, i.e., a straightforward comparison of $I(F_i;C)$ and $I(F_j;C)$ enables the determination of the feature that is potentially redundant and should be further analyzed. We note here that Theorem 2 is also the theoretical base for the methods such as FCBF (Yu and Liu, 2004) and Fast-FCBF (Song et al., 2013), which apply the approximate Markov blanket to identify redundancy.

5.3. Redundancy elimination

The feature remaining after the above two steps is a redundant candidate, which needs to be tested for significant redundancy. However, it is challenging to set a pervasive statistical test for redundancy analysis because the data distribution and the sample size vary depending on the conditions. Conventionally, the thresholding approach is widely applied to address this issue because of convenience. In this work, we apply a *top-down* approach for more reliable redundancy analysis. That is, we first sort the features selected by NOMP in the descending order of their class-relevance; then, we compare the pairwise correlations between features within a *top-down* order: According to Theorem 2, the features sorted in the top (i.e., with higher values of $I(F;C)$) are dominant compared with their followers. From the reliability perspective, the features sorted below are more likely to be significantly redundant. Therefore, we only need to test the redundancy for the features (F_j), which are sorted below the current one (F_i) and simultaneously satisfy the non-complementarity rule:

$$I(F_j;C) > I(F_j;C|F_i). \tag{29}$$

Then, if the correlation between F_i and F_j is larger than the class-relevance of the non-dominant feature F_i, F_j is identified as redundant and finally removed from the feature subset. Thus, we apply the following redundancy rule

$$I(F_j;C) < I(F_i;F_j) \tag{30}$$

to finally indicate that F_j is significantly redundant in the proposed method.

6. Proposed Method

The proposed method given in Algorithm 3, called feature selection with sparse representation and dependence analysis (SRDA), dynamically combines NOMP and dependence analysis. We also provide a toy example of an iteration of the proposed method on a dataset with 16 features, which is shown in Fig. 3.

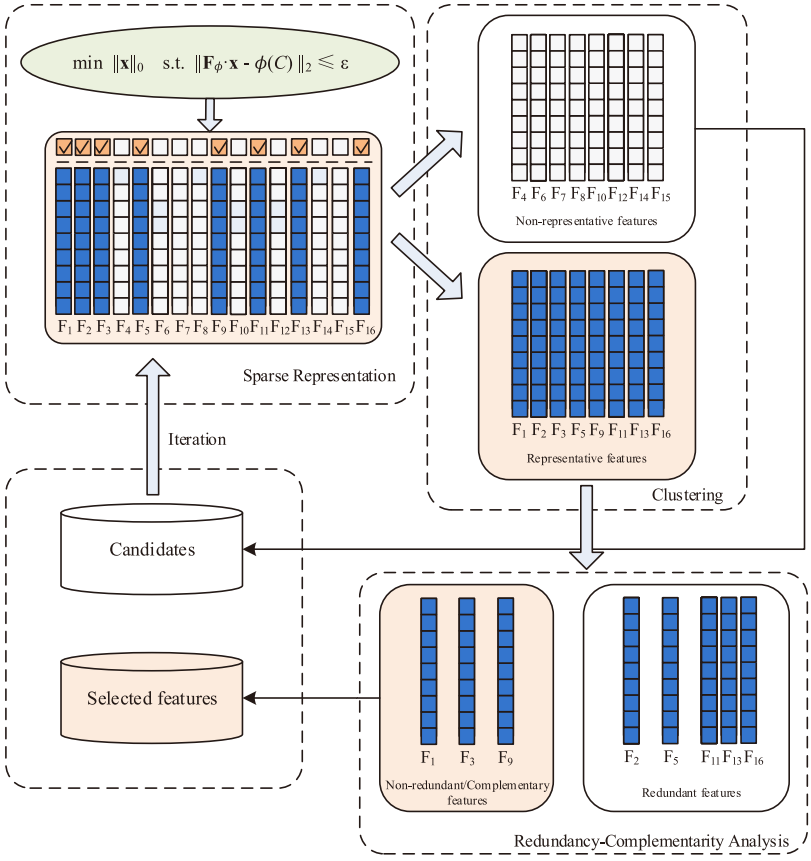

Figure 3: A toy example of one iteration of the proposed method.

In order to further utilize the high capability of MI in identifying relevant features, we marginally modify the NOMP shown in Algorithm 2 by using

$$\widetilde{F} = \arg \max_{F \in \mathbf{F} \setminus \mathbf{S}} I(F; C|\mathbf{S}) \tag{31}$$

to determine the initial feature for NOMP. However, the estimation of the joint distribution of the selected features in $\mathbf{S}$ is impeded by sample insufficiency. Thus, we apply the following approximation of Eq.(31):

$$\widetilde{F} = \arg \max_{F \in \mathbf{F} \setminus \mathbf{S}} \left\{ \min_{F' \in \mathbf{S}} I(F; C|F') \right\} \tag{32}$$

to determine the initial feature. Herein, $\min_{F' \in \mathbf{S}} I(F; C|F')$ is the pairwise pessimistic approximation of $I(F; C|\mathbf{S})$ (Wang et al., 2004).

Input: $\mathbf{F}$ /*feature set*/, C /*class*/, δ /*expected # features to be selected*/

Output: $\mathbf{S}$ /*selected feature subset*/, Ω /*Indices of selected features in $\mathbf{F}$*/

Initialize: Calculate Γ_ϕ and $\mathbf{\Lambda}_\phi$, $\mathbf{r}^0 \leftarrow C$, $\mathbf{S} \leftarrow \varnothing$, $\Omega^0 \leftarrow \varnothing$, $i \leftarrow 1$ /*iteration counter for NOMP*/

repeat

  $\mathbf{S}_{\text{NOMP}} \leftarrow \varnothing$, $\Omega^0 \leftarrow \varnothing$, $i \leftarrow 1$

  $\Omega^1 \leftarrow \Omega^0 \cup \{j | F_j = \arg\max_{F \in \mathbf{F} \setminus \mathbf{S}} \{\min_{F' \in \mathbf{S}} I(F; C|F')\}\}$

  repeat

    Find the feature $\widetilde{F}_j$ satisfying the maximization problem:

    $\max_{F \in \mathbf{F}} \{\Gamma_\phi(j) - \mathbf{\Lambda}_\phi(j, \Omega^i) \cdot \left(\mathbf{\Lambda}_\phi(\Omega^i, \Omega^i) + \lambda \cdot \mathbf{I}\right)^{-1} \cdot \Gamma_\phi^T(\Omega^i)\}$

    $\Omega^{i+1} = \Omega^i \cup \{j\}$

    Calculate $||\phi(\mathbf{r}^i)||_2 = \sqrt{\kappa(\mathbf{r}^i, \mathbf{r}^i)}$ in terms of Eq.(15)

    $i \leftarrow i + 1$

  until $\frac{||\phi(\mathbf{r}^i)||_2}{||\phi(\mathbf{r}^{i-1})||_2} \geq 0.95$;

  $\mathbf{S} \leftarrow \mathbf{S} \cup \mathbf{F}_{\Omega^{i+1}}$

  /* Dependence analysis step */

  $\mathbf{S}_{order} \leftarrow Sort(\mathbf{S}, \text{'descend'})$ /* sort the features in $\mathbf{S}$ in the descending order of $I(F_i; C)$*/

  $F_i \leftarrow \text{getFirstElement}(\mathbf{S}_{order})$

  repeat

    $F_j \leftarrow \text{getNextElement}(\mathbf{S}_{order}, F_i)$

    repeat

      if $I(F_j; C) > I(F_j; C|F_i)$ **and** $I(F_j; C) < I(F_i; F_j)$ **then**

        Remove F_j from $\mathbf{S}_{order}$

      end

      $F_j \leftarrow \text{getNextElement}(\mathbf{S}_{order}, F_j)$

    until $F_j == NULL$;

    $F_i \leftarrow \text{getNextElement}(\mathbf{S}_{order}, F_i)$

  until $F_i == NULL$;

  $\mathbf{S} \leftarrow \mathbf{S}_{order}$

until $|\mathbf{S}| \geq \delta$;

return $\mathbf{S}$

Algorithm 3: Feature selection with Sparse Representation and Dependence Analysis (SRDA)

The next section describes the extensive classification experiments conducted to empirically illustrate the effectiveness of the proposed method compared with the representative feature selection methods.

7. Experiments and Discussion

Four representative information theoretic feature selection methods, namely, MIM (Lewis, 1992), mRMR (Peng et al., 2005), FOU (Brown et al., 2012), and JMI (Meyer et al., 2008), and an $\ell_{2,p}$-norm regularized discriminative feature selection method (DFS, Tao et al., 2016) are used to compare with the proposed SRDA. Three representative classifiers, namely, k-nearest neighbor (kNN, Aha and Kibler, 1991), naïve Bayes classifier (NBC, Witten and Frank, 2000), and random forest (Breiman, 2001), are selected to generate the classification error rate on the datasets represented with the selected features, because of their proven effectiveness in real-world applications.

Weka (Waikato environment for knowledge analysis, Witten and Frank, 2000) is selected as the classification platform. Because MIM has already been integrated in Weka, we directly use it in Weka to generate datasets with its selected features prior to classification. mRMR, FOU, and JMI are implemented in Java and with Weka interfaces. DFS and the proposed SRDA are implemented in Matlab. Following Liang and Hu (2017) and Wang et al. (2017), we set $k = 1$ for kNN. For random forest, we use the default parameter setting in Weka. For DFS, we set $p = 1$ as suggested by Tao et al. (2016).

7.1. Datasets

Nine frequently-used datasets from different fields (such as industrial engineering, bioinformatics, and image processing) are applied in the experiments, wherein eight of them are randomly selected from the UCI Machine Learning Repository[1]. Since we want to verify the efficiency and effectiveness of the proposed method on high-dimensional data, we also select a well-known microarray dataset, 14_Tumors (Ramaswamy et al., 2001), as the benchmark dataset. General information about these datasets is summarized in Tab. 1.

Table 1: Description of datasets

#	Name	# samples	# features	# classes
1	isolet5	1559	617	26
2	DNA	3186	180	3
3	mfeat-factors	2000	216	10
4	mfeat-pixel	2000	240	10
5	mfeat-zernike	2000	47	10
6	optdigits	5620	63	10
7	spambase	4601	57	2
8	musk2	6598	166	2
9	14_Tumors	308	15009	26

[1] https://archive.ics.uci.edu

Isolet5 is a dataset for the prediction task of speech recognition. It contains 617 voice-related features and a total of 1559 samples. The class labels are the 26 letters in the alphabet. DNA dataset consisting of 3186 samples and 180 indicator binary features is created for recognition the boundaries between exons and introns given a sequence of DNA. The dataset applied in our experiments is slightly different from the original one in the UCI Machine Learning Repository (the original 60 symbolic attributes were changed into 180 binary attributes and four samples with ambiguities were removed)[2]. Multiple Features (mfeat) is a set of datasets that consist of features of handwritten numerals extracted from a collection of Dutch utility maps. In our experiments, three kinds of them, namely, mfeat-factors, mfeat-pixel, and mfeat-zernike, are selected as the benchmark datasets. Optical Recognition of Handwritten Digits (optdigits) dataset contains 5620 samples and 64 integer features in the range of [0,16]. The first feature of this dataset is removed in our experiments for its value never changes. Spambase contains a total of 4601 spam and non-spam email samples and 57 features indicating the length of sequences of consecutive capital letters and whether a particular word or character frequently occurs in the email. Musk2 dataset describes a set of 102 molecules of which 39 are judged by human experts to be musks and the remaining 63 molecules are judged to be non-musks. It contains 6598 samples and 166 features. 14_Tumors consists of 308 samples with a total of 26 classes including 14 various human tumor types (leukemia, prostate, lung, colorectal, lymphoma, bladder, melanoma, uterus, breast, renal, pancreas, ovary, mesothelioma, and CNS) and 12 normal tissues (breast, prostate, lung, colon, germinal center, bladder, uterus, peripheral blood, kidney, pancreas, ovary, and brain). Each sample has 15009 genes (features).

7.2. Experimental settings

First, we illustrate the classification results of the three classifiers on the top δ selected features for each feature selection method; here, δ is the desired number of selected features, specified as $\delta = [1, 2, ..., t]$. The maximal acceptable size t is set as min $\{100, |\mathbf{F}|\}$. The 5-fold cross-validation is performed to obtain the average classification error and the corresponding standard deviation. That is, each dataset is randomly partitioned into five complementary folds, wherein each fold as well as the other four folds will be represented using the features selected by feature selection methods working on the other four folds. The current fold (test set) will be then applied to validate the classification models trained on the other four folds (training set). After five rounds of the procedure mentioned above, the average classification result will be finally obtained and reported.

To further verify the superiority of the proposed method, we statistically compare the classification results on the datasets with an identical number of selected features for all the compared methods. Specifically, the top 20 and 40 features, respectively, are applied to obtain the classification results. Because sufficient

[2]https://www.rdocumentation.org/packages/mlbench/versions/2.1-1/topics/DNA

samples are required for the statistical test to obtain reliable results, the 5-fold cross-validation is repeated 20 times with different random seeds to generate 100 classification error samples. The average of these samples is reported, and the Wilcoxon rank-sum test with a significance level of 0.05 is applied to determine the statistical significance of the differences of such results.

7.3. Experimental results and discussion

Figures 4(a) – 4(i) and Figs. 5(a) – 5(i) show the error rates and the corresponding standard deviations of kNN and NBC, respectively, w.r.t the number of selected features on the nine datasets via 5-fold cross-validation. For 14_Tumors, the large standard deviation bar is likely to affect the comparison and thus is omitted in Figs. 4(i) and 5(i).

According to the results shown in Figs. 4(a) – 4(i), the superiority of SRDA can be verified in the majority of cases, particularly, on six datasets: isolet5 (Fig. 4(a)), mfeat-factors (Fig. 4(c)), mfeat-pixel (Fig. 4(d)), mfeat-zernike (Fig. 4(e)), musk2 (Fig. 4(h)), and 14_Tumors (Fig. 4(i)). It should be noted that although DFS is claimed in Tao et al. (2016) to be capable of addressing redundancy, our experimental results demonstrate that DFS performs marginally inferiorly on the whole, among the selected methods. This is possibly because DFS does not sufficiently consider redundancy and complementarity. Moreover, DFS is computationally intractable on 14_Tumors, which contains 15009 features. This empirically implies the limitations of the feature selection methods based on linear discriminative analysis.

However, for the rest datasets (i.e., DNA, optdigits, and spambase), all the selected methods except DFS (which performs inferiorly) exhibit similar performance, indicating the similar utilities of their selected features in data representation. It is also observed that the proposed SRDA performs relatively inferiorly at the top of the selected features (e.g., top 1–10 features of isolet5 and mfeat-pixel and top 1–30 features of mfeat-factors and 14_Tumors, respectively); this implies that redundancy is not fully eliminated among the top selected features. This phenomenon could be owing to the fact that SRDA conducts dependence analysis only locally (i.e., only considers redundancy and complementarity within the feature group generated by NOMP). On the whole, the proposed SRDA achieves no significantly inferior results on any of the selected datasets. The comparison results of NBC and random forest shown in Figs. 5(a) – 6(i) are similar to those of kNN, which also verify the effectiveness of SRDA.

Tables 2 – 4 present the classification error rate of kNN, NBC, and random forest over the 20×5-fold cross-validation on the top 20 and 40 selected features, respectively. For each dataset, the Wilcoxon test is conducted to evaluate the statistical significance of the difference between two sequences of the classification results, i.e., the sequence of the 20×5-fold cross-validation results corresponding to SRDA and that corresponding to any other feature selection method. The *Err* column records the average error rate of the 20×5-fold cross-validation. The *p-val* column records the p-value associated with the Wilcoxon test, where a p-value less than 0.05 indicates statistical significance. The notations $\bullet/\circ$ are used to illustrate that

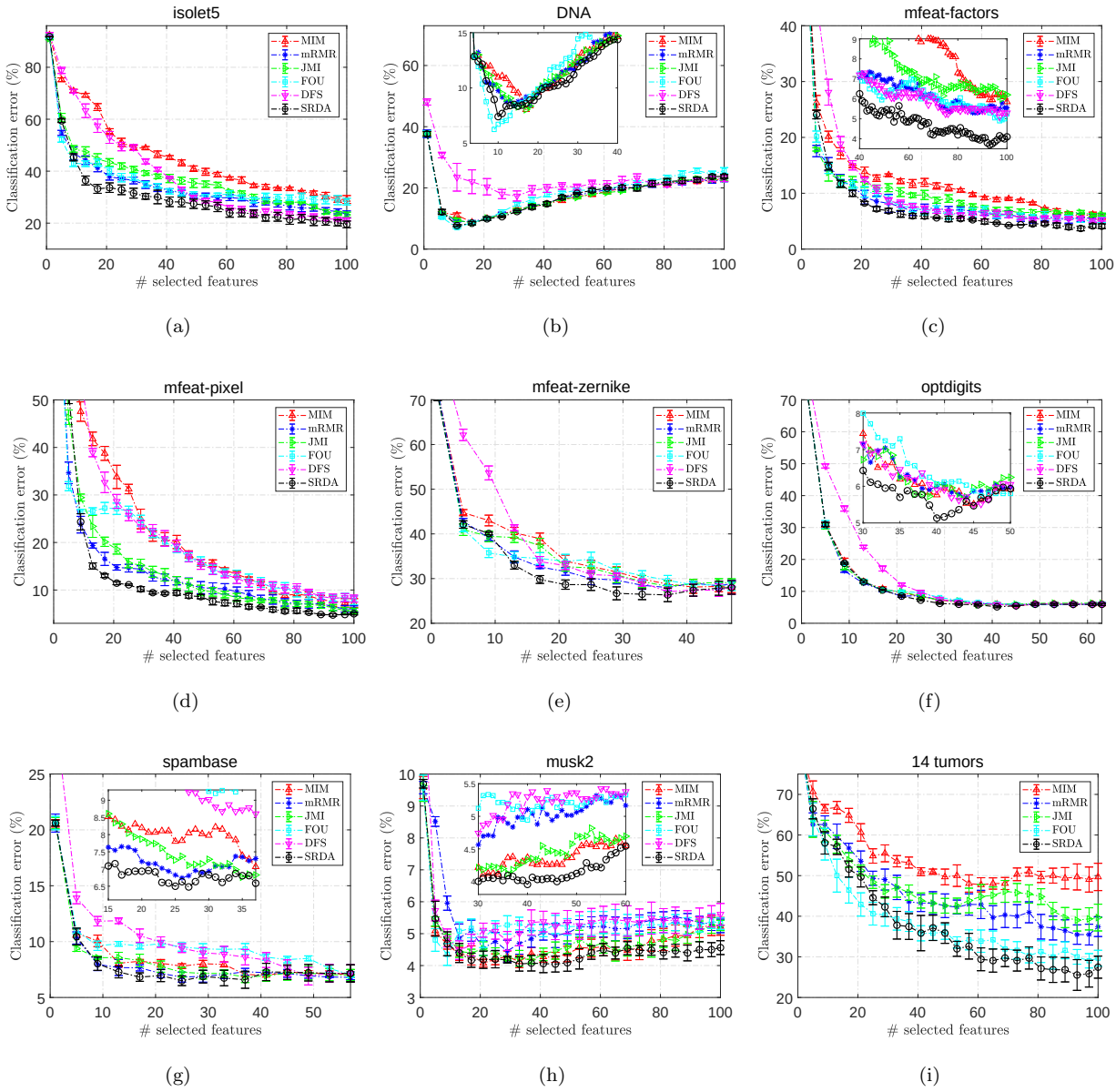

Figure 4: Accuracy comparison using kNN with different number of selected features on the selected datasets.

the average error rate corresponding to the current feature selection method is significantly lower/higher than that of SRDA. The bold value in each row represents the best classification result (i.e., the lowest error rate). The average error rate for all the datasets is reported in the last two rows. In addition, Tab. 5 reports the numbers of the significant losses/significant wins/ties of the selected methods compared with SRDA.

The results shown in Tabs. 2 – 4 illustrate that, SRDA achieves the best classification results in most of the cases. The average error rates for SRDA (18.12 (20 features)/15.61 (40 features) for kNN, 16.64 (20 features)/15.51 (40 features) for NBC, and 15.81 (20 features)/12.92 (40 features) for random forest) are

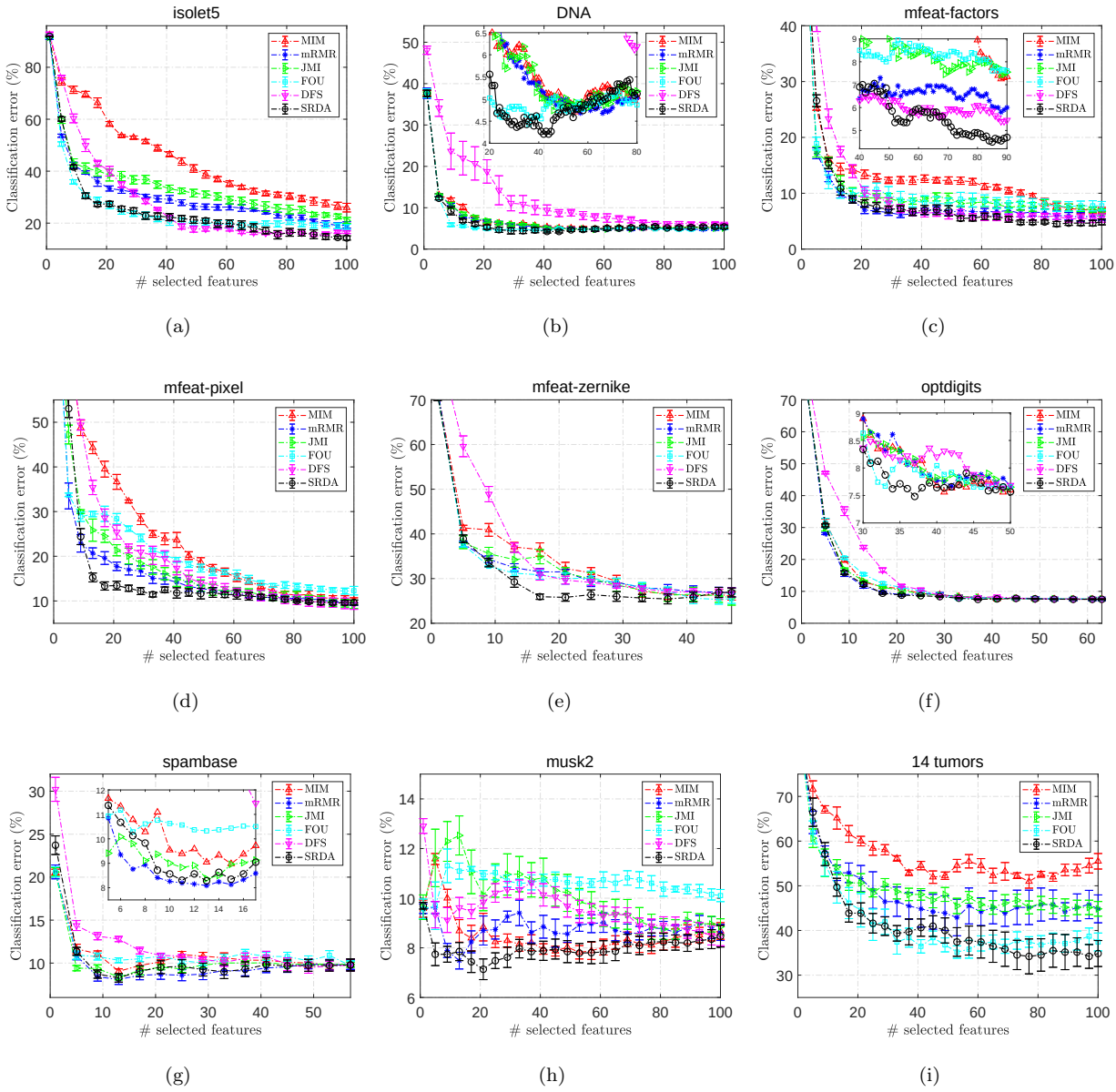

Figure 5: Accuracy comparison using NBC with different numbers of selected features on the selected datasets.

the lowest among all the feature selection methods with the three classifiers. In addition, the results of loss/win/tie shown in Tab. 5 also verify that SRDA significantly outperforms all the other methods. It can be also seen that MIM performs inferiorly to most of the selected methods especially on high-dimensional data, indicating the importance of feature inner-correlations for classification modeling. FOU, JMI, and mRMR explicitly conduct dependence analysis, so they perform significantly better than MIM and DFS.

We take an example on spambase dataset using MIM and DFS as the compared methods to further illustrate the characteristics of SRDA for supporting decision-making processes. The top five features selected by

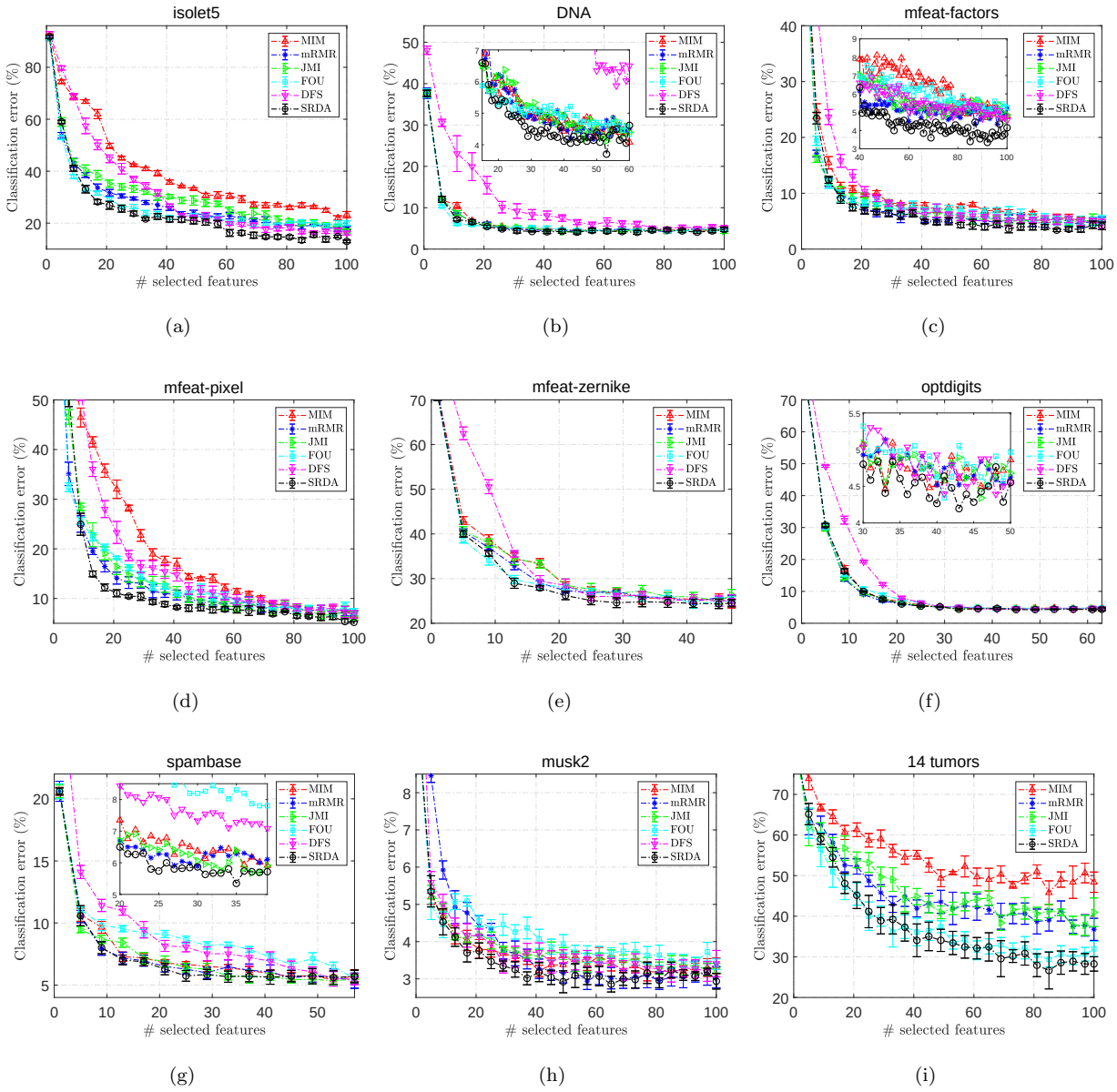

Figure 6: Accuracy comparison using random forest with different numbers of selected features on the selected datasets.

MIM and SRDA are "char_freq_!" (frequency of "!" in the email), "char_freq_$" (frequency of "$"), "capital_run_length_longest" (length of longest uninterrupted sequence of capital letters), "word_freq_remove" (frequency of "remove"), and "word_freq_your" (frequency of "your"). The results are interpretable in decision making processes because these words and items do frequently appear in junk emails like some advertisements for products/websites. However, the sixth selected feature of SRDA is "word_freq_free" (frequency of "free") while that of MIM is "capital_run_length_average" (average length of uninterrupted sequence of capital letters), where the former corresponds to better classification results for all the selected

Table 2: Results of classification error rate of kNN and Wilcoxon test for top 20 and 40 selected features.

# datasets	# features	SRDA	MIM		mRMR		FOU		JMI		DFS	
		Err	Err	p-val	Err	p-val	Err	p-val	Err	p-val	Err	p-val
1	20	**34.63**	56.12	0.000∘	39.35	0.000∘	40.14	0.000∘	43.04	0.000∘	56.74	0.000∘
	40	**29.20**	44.90	0.000∘	32.55	0.000∘	33.39	0.000∘	38.08	0.000∘	37.62	0.000∘
2	20	**9.30**	9.62	0.123	9.42	0.658	10.31	0.000∘	9.72	0.012∘	19.10	0.000∘
	40	**14.68**	14.77	0.581	14.93	0.158	17.01	0.000∘	14.58	0.808	19.44	0.000∘
3	20	**8.61**	14.22	0.000∘	9.70	0.000	9.89	0.000∘	11.66	0.000∘	12.52	0.000∘
	40	**6.25**	12.15	0.000∘	7.45	0.000∘	7.22	0.000∘	9.37	0.000∘	7.31	0.000∘
4	20	**12.06**	35.39	0.000∘	15.09	0.000∘	26.78	0.000∘	19.61	0.000∘	29.50	0.000∘
	40	**8.75**	20.02	0.000∘	11.35	0.000∘	17.97	0.000∘	12.62	0.000∘	19.66	0.000∘
5	20	**28.74**	35.88	0.000∘	31.23	0.000∘	34.48	0.000∘	34.80	0.000∘	31.13	0.000∘
	40	**27.40**	28.59	0.000	28.69	0.000	29.33	0.000∘	28.61	0.000∘	28.05	0.000∘
6	20	9.04	10.11	0.000∘	**8.95**	0.532	10.62	0.000∘	9.93	0.000∘	13.08	0.000∘
	40	**5.78**	6.11	0.000∘	6.11	0.001∘	6.32	0.000∘	6.11	0.000∘	6.00	0.010∘
7	20	**6.96**	7.89	0.000∘	7.45	0.000∘	9.53	0.000∘	7.56	0.000∘	9.95	0.000∘
	40	7.01	6.96	0.861	7.18	0.048∘	8.96	0.000∘	7.010	0.740	8.28	0.000∘
8	20	**4.08**	4.32	0.001∘	4.98	0.000∘	4.95	0.000∘	4.36	0.000∘	5.12	0.000∘
	40	**4.05**	4.22	0.014∘	5.09	0.000∘	5.15	0.000∘	4.45	0.000∘	5.45	0.000∘
9	20	49.68	59.50	0.000∘	53.87	0.005∘	**45.10**	0.000•	53.53	0.008	N/A	N/A∘
	40	**37.34**	51.65	0.000∘	45.00	0.000∘	38.13	0.442	46.02	0.000∘	N/A	N/A∘
Avg.	20	**18.12**	25.89		20.00		21.31		21.58		22.14*	
	40	**15.61**	21.04		17.59		18.16		18.54		16.48*	

∘ denotes statistical degradation at significance level of 0.05, and • denotes statistical improvement at significance level of 0.05.

* Calculated only on eight datasets.

classifiers. From our experience, word "free" is more likely to co-occur with dollar sign "$" and word "your" in junk mails. In addition, "capital_run_length_average" seems redundant when "capital_run_length_longest" has been already selected, although both of them seem relevant to spam detection. This indicates the superiority of SRDA in redundancy and complementarity analysis. As for DFS, the top-ranked features performs significantly inferiorly in spam detection, indicating that redundancy is not fully removed. Worse still, the features selected by DFS, e.g., "char_freq_(" (frequency of char "(") in the top-ranked features, are less interpretable for spam analysis and detection.

Table 3: Results of classification error rate of NBC and Wilcoxon test for top 20 and 40 selected features.

# datasets	# features	SRDA	MIM		mRMR		FOU		JMI		DFS	
		Err	Err	p-val	Err	p-val	Err	p-val	Err	p-val	Err	p-val
1	20	27.75	58.73	0.000∘	34.31	0.000∘	**26.11**	0.000•	39.33	0.000∘	41.46	0.000∘
	40	23.22	46.52	0.000∘	28.28	0.000∘	**20.15**	0.000•	34.55	0.000∘	22.98	0.723
2	20	5.54	6.80	0.000∘	6.79	0.000∘	**4.96**	0.000•	6.72	0.000∘	16.48	0.000∘
	40	**4.20**	5.27	0.000∘	5.19	0.000∘	4.84	0.000∘	5.26	0.000∘	8.75	0.000∘
3	20	**8.02**	13.48	0.000∘	8.10	0.485	9.56	0.000∘	10.12	0.000∘	10.75	0.000∘
	40	7.06	12.28	0.000∘	**6.77**	0.098	8.35	0.000∘	9.27	0.000∘	7.18	0.899
4	20	**13.19**	36.88	0.000∘	18.38	0.000∘	29.21	0.000∘	21.19	0.000∘	26.12	0.000∘
	40	**11.55**	22.44	0.000∘	14.11	0.000∘	19.44	0.000∘	15.40	0.000∘	18.01	0.000∘
5	20	**25.75**	34.07	0.000∘	29.95	0.000∘	30.89	0.000∘	33.45	0.000∘	28.21	0.000∘
	40	**25.71**	27.00	0.000∘	27.05	0.000∘	27.33	0.000∘	27.07	0.000∘	26.21	0.111
6	20	**9.01**	11.11	0.000∘	9.25	0.029∘	11.43	0.000∘	10.80	0.000∘	12.10	0.000∘
	40	**7.74**	7.83	0.438	7.75	0.961	8.04	0.011∘	7.78	0.737	8.28	0.000∘
7	20	9.35	10.08	0.000∘	**8.76**	0.000•	10.76	0.000∘	9.42	0.842	10.98	0.000∘
	40	9.79	10.26	0.001∘	**9.16**	0.000•	10.89	0.000∘	10.06	0.033∘	10.41	0.000∘
8	20	**7.36**	8.36	0.000∘	8.20	0.000∘	11.03	0.000∘	10.64	0.000∘	9.90	0.000∘
	40	**7.71**	7.97	0.004∘	8.73	0.000∘	10.65	0.000∘	10.69	0.000∘	10.18	0.000∘
9	20	**43.83**	57.34	0.000∘	50.42	0.000∘	44.08	0.224	49.92	0.003∘	N/A	N/A∘
	40	42.58	52.05	0.000∘	45.24	0.002∘	**39.29**	0.000•	46.97	0.000∘	N/A	N/A∘
Avg.	20	**16.64**	26.32		19.35		19.78		21.29		19.50*	
	40	**15.51**	21.29		16.92		16.55		18.56		14.00*	

∘ denotes statistical degradation at significance level of 0.05, and • denotes statistical improvement at significance level of 0.05.
* Calculated only on eight datasets.

7.4. Runtime comparison for DFS and SRDA

In order to show the relative efficiency of SRDA, we compare the runtime of SRDA and DFS on DNA, isolet5, mfeat-factors, mfeat-pixel, mfeat-zernike, musk2, optdigits, and spambase datasets, respectively. The results are given in Fig. 7. For each dataset, we report the runtime of the methods when they select 100 features. Because DFS is computationally intractable on 14_Tumors within reasonable time frame, we do not show the results of either methods on 14_Tumors.

As can be seen from Fig. 7, SRDA runs significantly faster than DFS on all the selected datasets, particularly on isolet5 (containing 1559 samples and 617 features), mfeat-factors (containing 2000 samples and 216 features), and mfeat-pixel (containing 2000 samples and 240 features), verifying the efficiency of the pro-

Table 4: Results of classification error rate of random forest and Wilcoxon test for top 20 and 40 selected features.

# datasets	# features	SRDA	MIM		mRMR		FOU		JMI		DFS	
		Err	Err	p-val	Err	p-val	Err	p-val	Err	p-val	Err	p-val
1	20	**28.33**	51.01	0.000∘	32.21	0.000∘	29.00	0.040∘	35.70	0.000∘	47.79	0.000∘
	40	**22.04**	37.21	0.000∘	25.19	0.000∘	22.89	0.007∘	30.57	0.000∘	27.74	0.000∘
2	20	5.63	5.97	0.004∘	5.80	0.049∘	**5.56**	0.866	6.03	0.008∘	14.40	0.000∘
	40	**4.18**	4.64	0.000∘	4.63	0.000∘	4.81	0.000∘	4.63	0.000∘	7.05	0.000∘
3	20	**7.12**	9.86	0.000∘	7.80	0.000∘	8.51	0.000∘	8.74	0.000∘	10.73	0.000∘
	40	5.88	7.76	0.000∘	**5.60**	0.098	6.92	0.000∘	6.59	0.000∘	6.98	0.826
4	20	**11.91**	32.37	0.000∘	14.63	0.000∘	18.50	0.000∘	17.50	0.000∘	23.18	0.000∘
	40	**8.56**	16.77	0.000∘	10.39	0.000∘	11.15	0.000∘	11.67	0.000∘	14.57	0.000∘
5	20	**27.13**	30.88	0.000∘	28.19	0.000∘	28.75	0.000∘	30.49	0.000	27.35	0.536
	40	25.38	25.45	0.929	25.52	0.623	26.34	0.000∘	25.48	0.656	**25.31**	0.816
6	20	**6.09**	6.89	0.000∘	6.32	0.000∘	7.72	0.000∘	6.77	0.000∘	8.30	0.000∘
	40	4.63	4.72	0.244	4.62	0.831	4.60	0.638	4.70	0.458	4.72	0.591∘
7	20	**6.46**	6.83	0.001∘	6.63	0.142	8.86	0.000∘	6.69	0.034∘	8.49	0.000∘
	40	5.60	**5.57**	0.792	5.99	0.000∘	7.36	0.000∘	5.64	0.406	6.68	0.000∘
8	20	**3.83**	3.97	0.100	4.58	0.000∘	4.58	0.000∘	4.04	0.008∘	4.36	0.000∘
	40	**3.34**	3.43	0.195	3.47	0.067	3.98	0.000∘	3.70	0.000∘	3.86	0.000∘
9	20	**45.77**	59.82	0.000∘	52.60	0.000∘	45.81	0.866	53.05	0.000∘	N/A	N/A∘
	40	**36.69**	51.37	0.000∘	44.79	0.000∘	39.11	0.044∘	46.24	0.000∘	N/A	N/A∘
Avg.	20	**15.81**	23.07		17.64		17.48		18.78		18.08*	
	40	**12.92**	17.44		14.47		14.13		15.47		12.11*	

∘ denotes statistical degradation at significance level of 0.05, and • denotes statistical improvement at significance level of 0.05.

* Calculated only on eight datasets.

posed method. In addition, the results also show that DFS is much more sensitive to the number of features and requires a significantly longer runtime when the feature size increases. This as well as the computational intractability on 14_Tumors both imply a dilemma that DFS and similar sparse representation-based feature selection methods will face when dealing with large-scale datasets.

8. Conclusions

In this paper, a novel feature selection method is proposed to discriminate the salient features for classification utilizing both sparse representation and information theoretic dependence analysis. Specifically, each

Table 5: Results of loss/win/tie.

	# features	MIM	mRMR	FOU	JMI	DFS
kNN	20	8/0/1	6/0/3	8/1/0	8/0/1	9/0/0
	40	6/0/3	7/0/2	8/0/1	7/0/2	9/0/0
NBC	20	9/0/0	7/1/1	6/2/1	8/0/1	9/0/0
	40	8/0/1	6/1/2	7/2/0	8/0/1	6/0/3
random forest	20	8/0/1	8/0/1	7/0/2	8/0/1	8/0/1
	40	5/0/4	5/0/4	8/0/1	6/0/3	7/0/2

The results are collected according to the p-val of the Wilcoxon test presented in Tabs. 2 and 4.

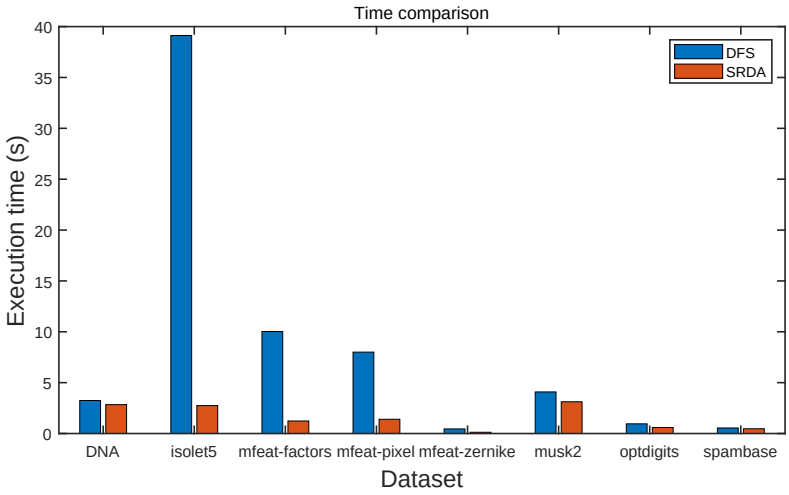

Figure 7: Execution time for DFS and SRDA on the selected datasets. The results for 14_Tumors are omitted because DFS is computationally intractable within reasonable time frame on 14_Tumors.

iteration of the proposed method first applies a nonlinear sparse representation approach to determine the representative feature cluster and then conducts approximate dependence analysis to eliminate redundancy in such a manner as to obtain the final selected features. This searching strategy can effectively prevent the significant bias caused by pairwise correlation analysis of large-scale feature set because the size of the selected feature cluster from each iteration is generally small owing to the effective sparse representation of features. For dependence analysis, the complementary correlation of features is first examined, and then, dominance analysis is conducted on the features with non-complementarity. Finally, redundancy analysis is conducted on the non-dominant features, which are finally eliminated if the redundancy rule is satisfied. The proposed method not only utilizes the optimization framework of sparse representation to select features, but also conducts dependence analysis to explicitly handle redundancy and complementarity. In addition,

explicit redundancy and complementarity analysis can improve interpretability in decision support processes. Extensive classification experiments are conducted for the proposed method and five representative feature selection methods, including four representative information theoretic feature selection methods and an $\ell_{2,p}$-norm regularized discriminative feature selection method. The experimental results on nine popular datasets verify the effectiveness and superiority of the proposed method.

However, our method exhibits certain evident limitations. For example, classes are always discrete whereas features are sometimes continuous, and we do not consider an effective approach toward measuring the correlation among different types of variables. Another limitation of this study is that our method still applies pairwise analysis as an approximation for the analysis of higher-order correlations, although the pairwise analysis of a not-so-large feature cluster can partially prevent estimation bias in the proposed method. Extending our method along such lines would require modeling a complex k-wise analytical framework that attempts to prevent estimation bias caused by the sample insufficiency—an analytically challenging task. Future research could profitably consider these and other extensions such as applying ℓ_p-norm $(0 < p < 1)$ which exhibits more desirable characteristics in sparse representation. In addition, the idea of this paper can be applied to an unsupervised feature selection scheme: Unsupervised dimensionality reduction approaches like principal components analysis (PCA) and information theoretic dependence analysis can be possibly integrated into a transfer learning method, yielding another promising future research direction.

Acknowledgement

We would like to thank the associate editor and three anonymous reviewers for their constructive comments and suggestions. This study was supported in part by the National Natural Science Foundation of China under Grants 71702066, 71802192, 61703319, and 71772077, in part by China Postdoctoral Science Foundation under Grant 2017M612856, in part by Humanity and Social Science Youth foundation of Ministry of Education of China under Grant 18YJC630137, and in part by National Key R&D Program of China under Grant 2017YFB0102500.

References

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

Breiman, L., 2001. Random forests. Machine Learning 45 (1), 5–32.

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.

Cai, T. T., Wang, L., 2011. Orthogonal matching pursuit for sparse signal recovery with noise. IEEE Transactions on Information Theory 57 (7), 4680–4688.

Candes, E. J., Tao, T., 2005. Decoding by linear programming. IEEE Transactions on Information Theory 51, 4203–4215.

Chen, X., Ge, D., Wang, Z., Ye, Y., 2014. Complexity of unconstrained ℓ_2-ℓ_p minimization. Mathematical Programming 143 (1-2), 371–383.

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.

Chen, Z., Wu, C., Zhang, Y., Huang, Z., Ran, B., Zhong, M., Lyu, N., 2015. Feature selection with redundancy-complementariness dispersion. Knowledge-Based Systems 89, 203–217.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Annals of Statistics 32 (2), 407–499.

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

Foucart, S., Lai, M.-J., 2009. Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q < 1$. Applied and Computational Harmonic Analysis 26 (3), 395–407.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and features election. Journal of Machine Learning Research 3, 1157–1182.

Huang, D., Chow, T. W. S., 2005. Effective feature selection scheme using mutual information. Neurocomputing 63, 325–343.

Kohavi, R., John, G. H., 1997. Wrappers for feature subset selection. Artificial Intelligence 97, 273–324.

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.

Liang, M., Hu, X., 2017. An efficient semi-supervised representatives feature selection algorithm based on information theory. Pattern Recognition 61, 511–523.

Liu, M., Zhang, D., 2016. Pairwise constraint-guided sparse learning for feature selection. IEEE Transactions on Cybernetics 46 (1), 298–310.

Luo, S., Chen, Z., 2014. Sequential lasso cum EBIC for feature selection with ultra-high dimensional feature space. Journal of the American Statistical Association 109, 1229–1240.

Maldonado, S., Bravo, C., López, J., Pérez, J., 2017. Integrated framework for profit-based feature selection and svm classification in credit scoring. Decision Support Systems 104, 113–121.

Masaeli, M., Fung, G., Dy, J. G., 2010. From transformation-based dimensionality reduction to feature selection. In: Proceedings of the 27th International Conference on International Conference on Machine Learning. ICML'10. Omnipress, USA, pp. 751–758.
URL http://dl.acm.org/citation.cfm?id=3104322.3104418

Meyer, P., Bontempi, G., 2006. On the use of variable complementarity for feature selection in cancer classification. Evolutionary Computation and Machine Learning in Bioinformatics 3907, 91–102.

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.

Mo, X., Monga, V., Bala, R., Fan, Z. G., 2014. Adaptive sparse representations for video anomaly detection. IEEE Transactions on Circuits and Systems for Video Technology 24 (4), 631–645.

Neoh, S. C., Zhang, L., Mistry, K., Hossain, M. A., Lim, C. P., Aslam, N., Kinghorn, P., sep 2015. Intelligent facial emotion recognition using a layered encoding cascade optimization model. Applied Soft Computing 34, 72–93.
URL http://dx.doi.org/10.1016/j.asoc.2015.05.006 https://linkinghub.elsevier.com/retrieve/pii/S1568494615003063

Nie, F., Huang, H., Cai, X., Ding, C. H., 2010. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. Advances in Neural Information Processing Systems 23, 1813–1821.

Pandit, D., Zhang, L., Chattopadhyay, S., Lim, C. P., Liu, C., sep 2018. A scattering and repulsive swarm intelligence algorithm for solving global optimization problems. Knowledge-Based Systems 156, 12–42.
URL https://linkinghub.elsevier.com/retrieve/pii/S0950705118302120

Peng, B., Wang, L., Wu, Y., 2016. An error bound for ℓ_1-norm support vector machine coefficients in ultra-high dimension. The Journal of Machine Learning Research 17 (1), 8279–8304.

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.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C. H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J. P., Poggio, T., Gerald, W., Loda, M., Lander, E. S., Golub, T. R., 2001. Multiclass cancer diagnosis using tumor gene expression signatures. Proceedings of the National Academy of Sciences of the United States of America 98 (26), 15149–15154.

Rodriguez-Lujan, I., Huerta, R., Elkan, C., Cruz, C. S., 2010. Quadratic programming feature selection. Journal of Machine Learning Research 11, 1491–1516.

Serrano-Silva, Y. O., Villuendas-Rey, Y., Yáñez-Márquez, C., 2018. Automatic feature weighting for improving financial decision support systems. Decision Support Systems 107, 78–87.

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.

Srisukkham, W., Zhang, L., Neoh, S. C., Todryk, S., Lim, C. P., jul 2017. Intelligent leukaemia diagnosis with bare-bones pso based feature optimization. Applied Soft Computing 56, 405–419.
URL http://dx.doi.org/10.1016/j.asoc.2017.03.024 https://linkinghub.elsevier.com/retrieve/pii/S1568494617301485

Sun, X., Liu, Y., Xu, M., Chen, H., Han, J., Wang, K., 2013. Feature selection using dynamic weights for classification. Knowledge-Based Systems 37, 541–549.

Tan, T. Y., Zhang, L., Neoh, S. C., Lim, C. P., oct 2018. Intelligent skin cancer detection using enhanced particle swarm optimization. Knowledge-Based Systems 158, 118–135.
URL https://doi.org/10.1016/j.knosys.2018.05.042 https://linkinghub.elsevier.com/retrieve/pii/S0950705118302879

Tao, H., Hou, C., Nie, F., Jiao, Y., Yi, D., 2016. Effective discriminative feature selection with nontrivial solution. IEEE Transactions on Neural Networks & Learning Systems 27 (4), 796–808.

Thi, H. A. L., Dinh, T. P., Le, H. M., Vo, X. T., 2015. DC approximation approaches for sparse optimization. European Journal of Operational Research 244 (1), 26–46.

Wang, G., Lochovsky, F. H., Yang, Q., 2004. Feature selection with conditional mutual information maximin in text categorization. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management. CIKM'04. ACM Press, New York, NY, USA, pp. 342–349.

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.

Wang, K., He, R., Wang, L., Wang, W., Tan, T., 2016. Joint feature selection and subspace learning for cross-modal retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence 38 (10), 2010–2023.

Wang, L., Chen, S., Wang, Y., 2010. A unified algorithm for mixed $\ell_{2,p}$-minimizations and its application in feature selection. Advances in Neural Information Processing Systems 23, 1813–1821.

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

Xu, Z., Chang, X., Xu, F., Zhang, H., 2012. $l_{1/2}$ regularization: A thresholding representation theory and a fast solver. IEEE Transactions on Neural Networks and Learning Systems 23 (7), 1013–1027.

Xu, Z., Zhang, H., Wang, Y., Chang, X., Liang, Y., 2010. $l_{1/2}$ regularization. Science China 53 (6), 1159–1169.

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

Zhang, L., Mistry, K., Jiang, M., Chin Neoh, S., Hossain, M. A., nov 2015a. Adaptive facial point detection and emotion recognition for a humanoid robot. Computer Vision and Image Understanding 140, 93–114.
URL http://dx.doi.org/10.1016/j.cviu.2015.07.007 https://linkinghub.elsevier.com/retrieve/pii/S1077314215001605

Zhang, L., Mistry, K., Lim, C. P., Neoh, S. C., feb 2018a. Feature selection using firefly optimization for classification and regression models. Decision Support Systems 106, 64–85.

Zhang, L., Mistry, K., Neoh, S. C., Lim, C. P., nov 2016. Intelligent facial emotion recognition using moth-firefly optimization. Knowledge-Based Systems 111, 248–267.
URL http://dx.doi.org/10.1016/j.knosys.2016.08.018 https://linkinghub.elsevier.com/retrieve/pii/S0950705116302799

Zhang, L., Srisukkham, W., Neoh, S. C., Lim, C. P., Pandit, D., mar 2018b. Classifier ensemble reduction using a modified firefly algorithm: An empirical evaluation. Expert Systems with Applications 93, 395–422.
URL https://linkinghub.elsevier.com/retrieve/pii/S0957417417306759

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

Zhang, Y., Yang, C., Yang, A., Xiong, C., Zhou, X., Zhang, Z., 2015b. Feature selection for classification with class-separability strategy and data envelopment analysis. Neurocomputing 166, 172 – 184.
URL http://www.sciencedirect.com/science/article/pii/S0925231215004609

Biography

Yishi Zhang received his B.S. degree in computer science from University of Electronic Science and Technology of China in 2009, and M.S. and Ph.D. degrees in software engineering and management science and engineering from Huazhong University of Science and Technology in 2011 and 2016, respectively. He is now a research associate at School of Management, Jinan University, and a post-doc at Joseph M. Katz Graduate School of Business at University of Pittsburgh. His research interests involve dimensionality reduction, topic modeling, and business intelligence.

Qi Zhang received his B.S. degree in mechanical and electrical integration, M.S. degree in Technological Economics and Management, and Ph.D. degree in Management Science and Engineering from Wuhan University of Technology, respectively. She is now an associate professor at School of Economics and Management, China University of Geosciences (Wuhan). Her research interests involve business intelligence and big data analytics in thefieldof digital business.

Zhijun Chen received his B.S. degree in mechanical engineering and automation from Wuhan University of Technology in 2009, and M.S. and Ph.D. degrees in transportation engineering and automotive engineering from Wuhan University of Technology, in 2012 and 2016, respectively. His research interests involve traffic safety, vehicle behavior recognition, machine learning, and big data analytics in intelligent transportation systems.

Jennifer Shang is a Full professor in the area of Business Analytics in Katz Graduate School of Business at University of Pittsburgh. She received her Ph.D. in Operations Management from University of Texas at Austin, a Master Degree from University of Iowa, and a Bachelor Degree in International Business from National Taiwan University.She has published in various journals, including Management Science, Information Systems Research, Marketing Science, European Journal of Operational Research, among others. She has won the EMBA Distinguished Teaching Award and several Excellence-in-Teaching Awards from the MBA/EMBA programs at Katz Business School.

Haiying Wei received her B.S. degree in Statistics from Renmin University of China in 1986, M.S. degree in Finance from Jinan University in 1997, and her Ph.D. degree in Management Science and Engineering from Huazhong University of Science and Technology in 2006. At present, she is a professor at School of Management, Jinan University, and serves as an executive member of China Institutions for Higher Learning. Her research interests involve statistical analysis and its application in marketing.

Highlights

- A nonlinear sparse representation method is applied to find salient feature clusters.
- An approximate feature dependence analysis strategy is proposed.
- Salient and interpretable features can be obtained by the proposed method.

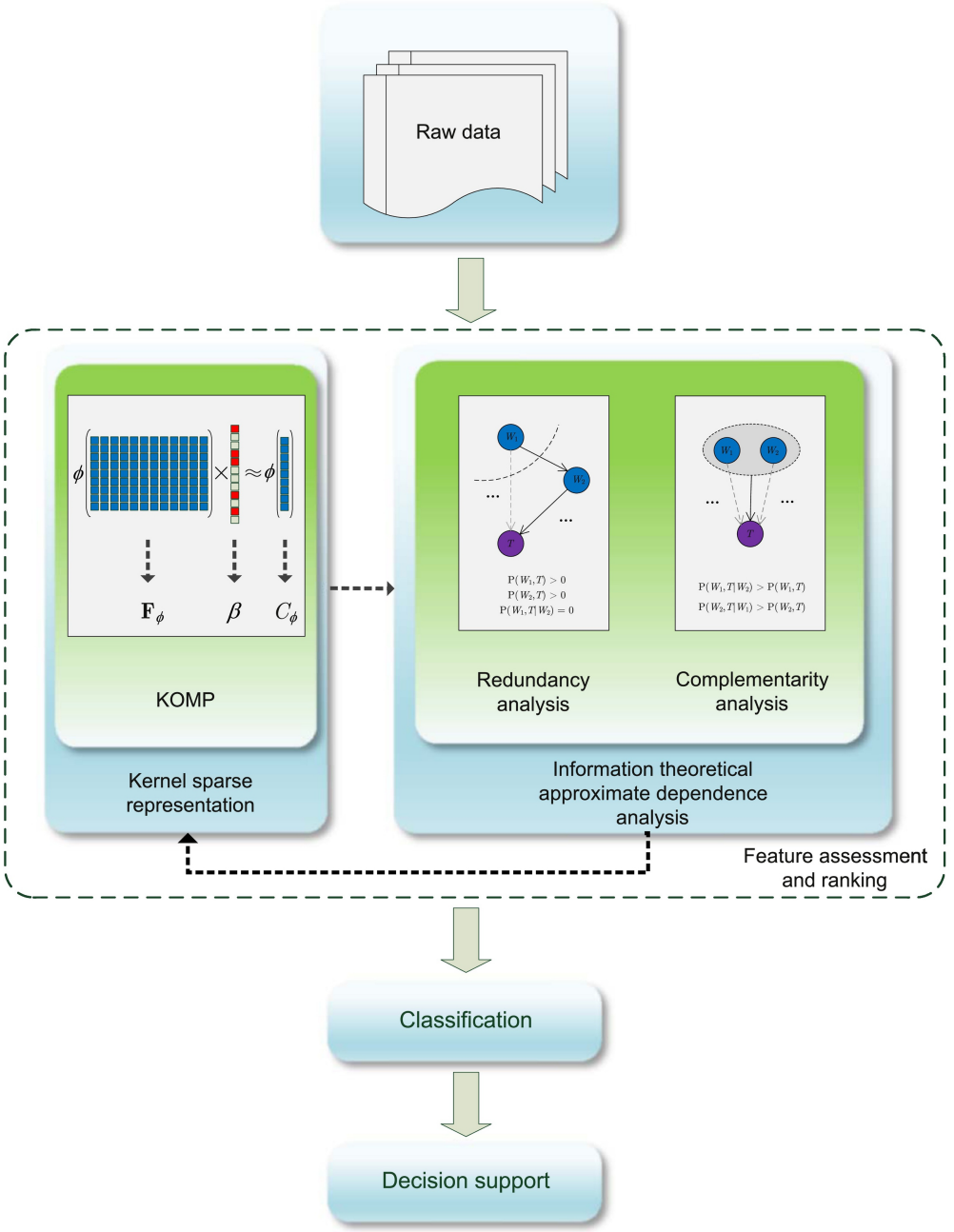
Raw data
KOMP
Kernel sparse representation
Redundancy analysis
Complementarity analysis
Information theoretical approximate dependence analysis
Feature assessment and ranking
Classification
Decision support

Figure 1

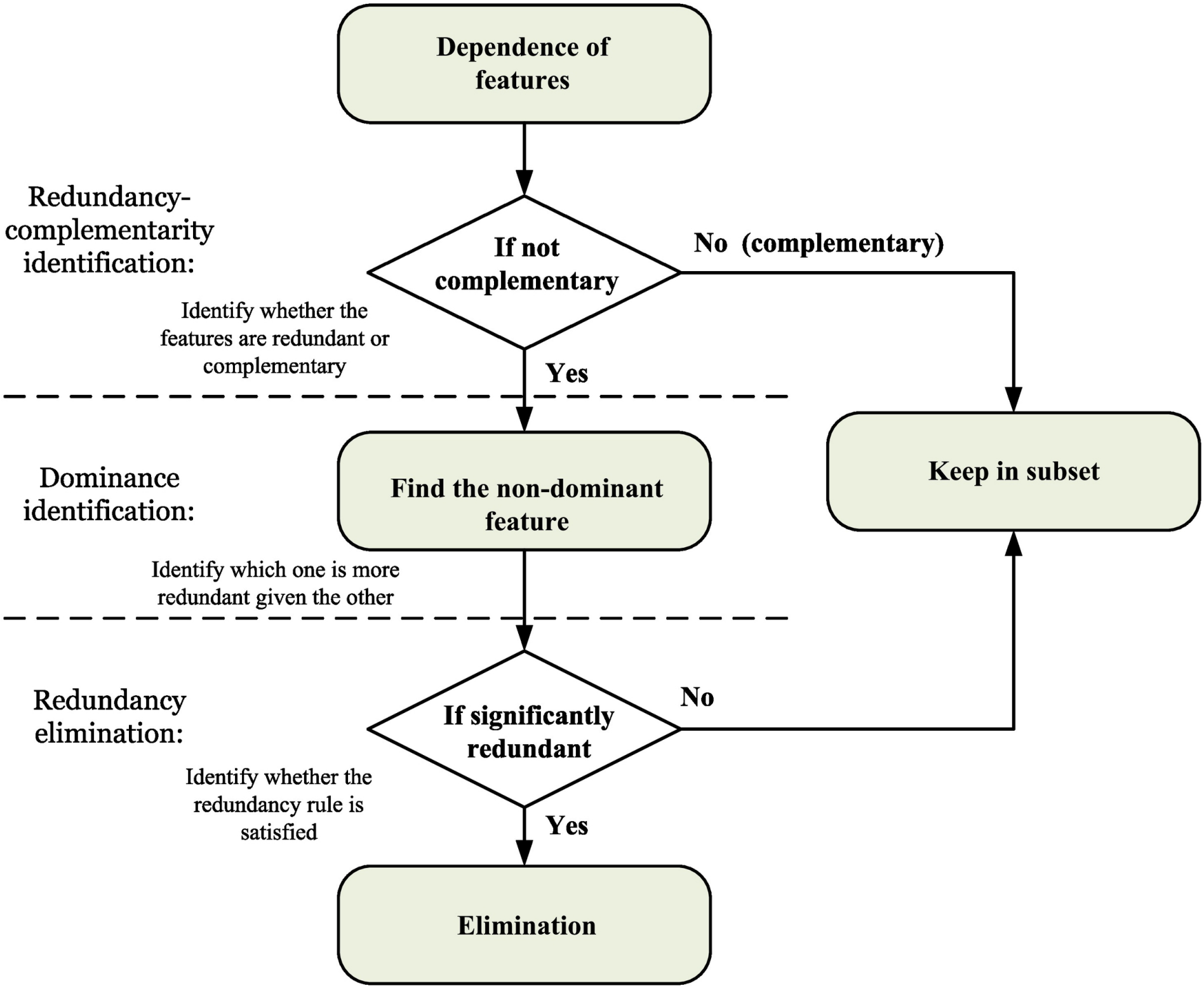
Dependence of features
Redundancy-complementarity identification:
If not complementary
No (complementary)
Identify whether the features are redundant or complementary
Yes
Keep in subset
Dominance identification:
Find the non-dominant feature
Identify which one is more redundant given the other
Redundancy elimination:
If significantly redundant
No
Identify whether the redundancy rule is satisfied
Yes
Elimination

Figure 2

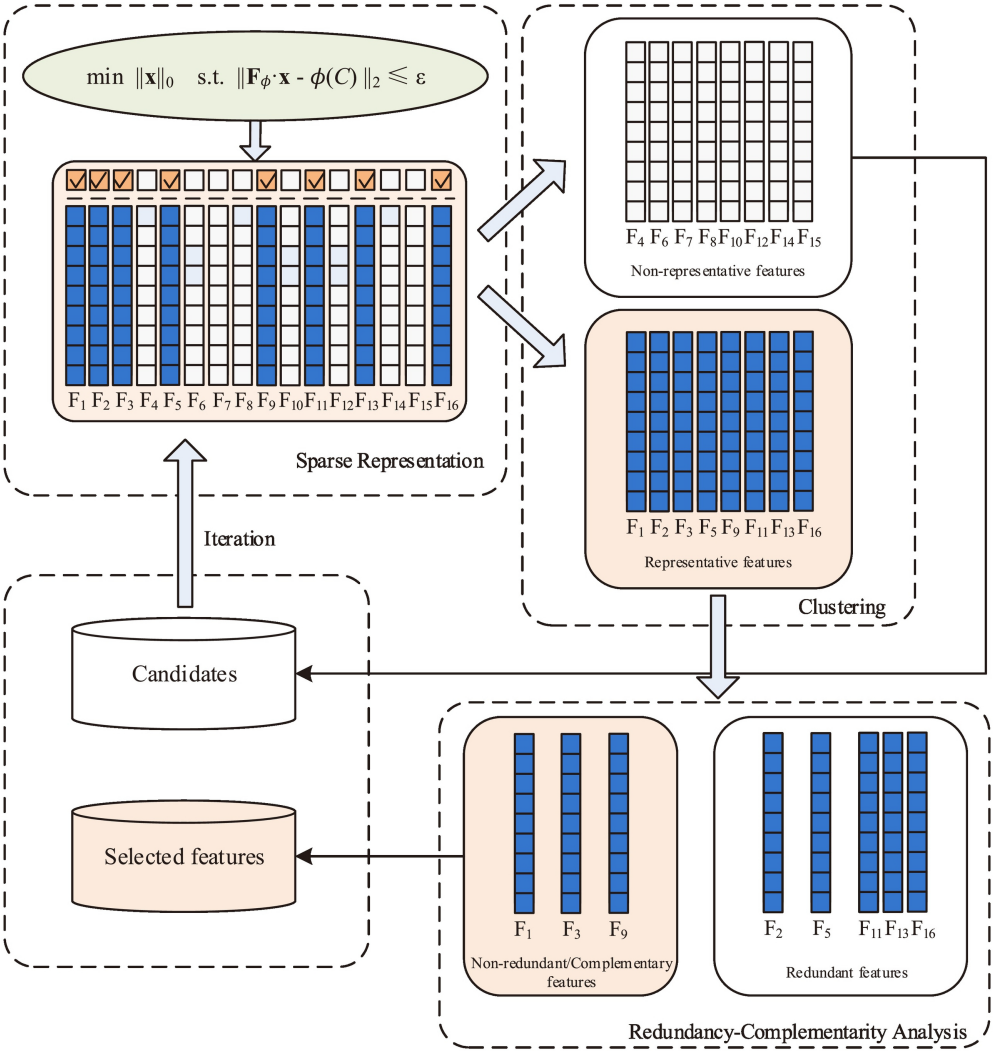
min ‖x‖0 s.t. ‖Fϕ·x - ϕ(C)‖2 ≤ ε
F1 F2 F3 F4 F5 F6 F7 F8 F9 F10 F11 F12 F13 F14 F15 F16
Sparse Representation
F4 F6 F7 F8 F10 F12 F14 F15
Non-representative features
F1 F2 F3 F5 F9 F11 F13 F16
Representative features
Clustering
Iteration
Candidates
Selected features
F1 F3 F9
Non-redundant/Complementary features
F2 F5 F11 F13 F16
Redundant features
Redundancy-Complementarity Analysis

Figure 3

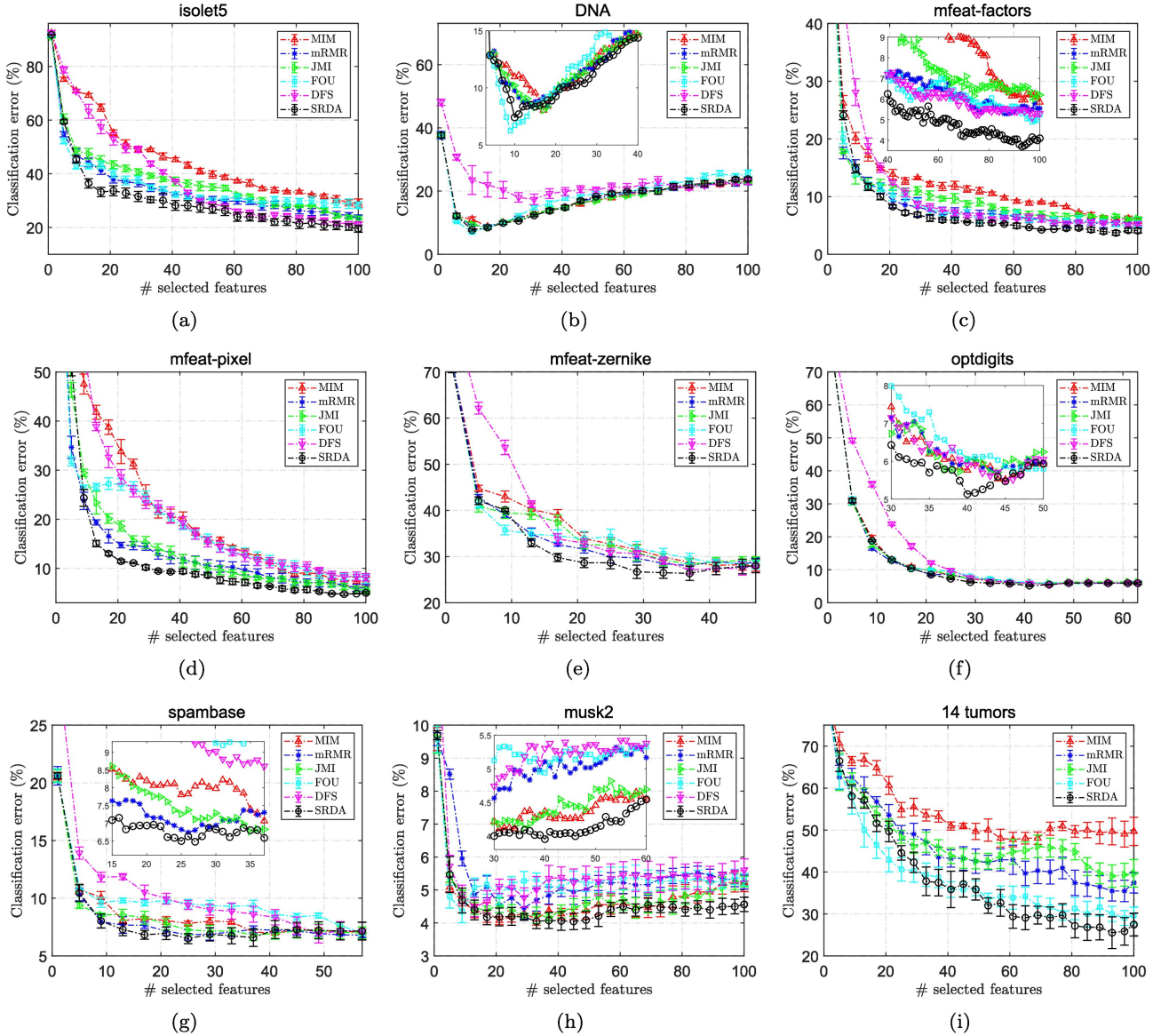
isolet5
DNA
mfeat-factors
mfeat-pixel
mfeat-zernike
optdigits
spambase
musk2
14 tumors
MIM
mRMR
JMI
FOU
DFS
SRDA
Classification error (%)
selected features
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)

Figure 4

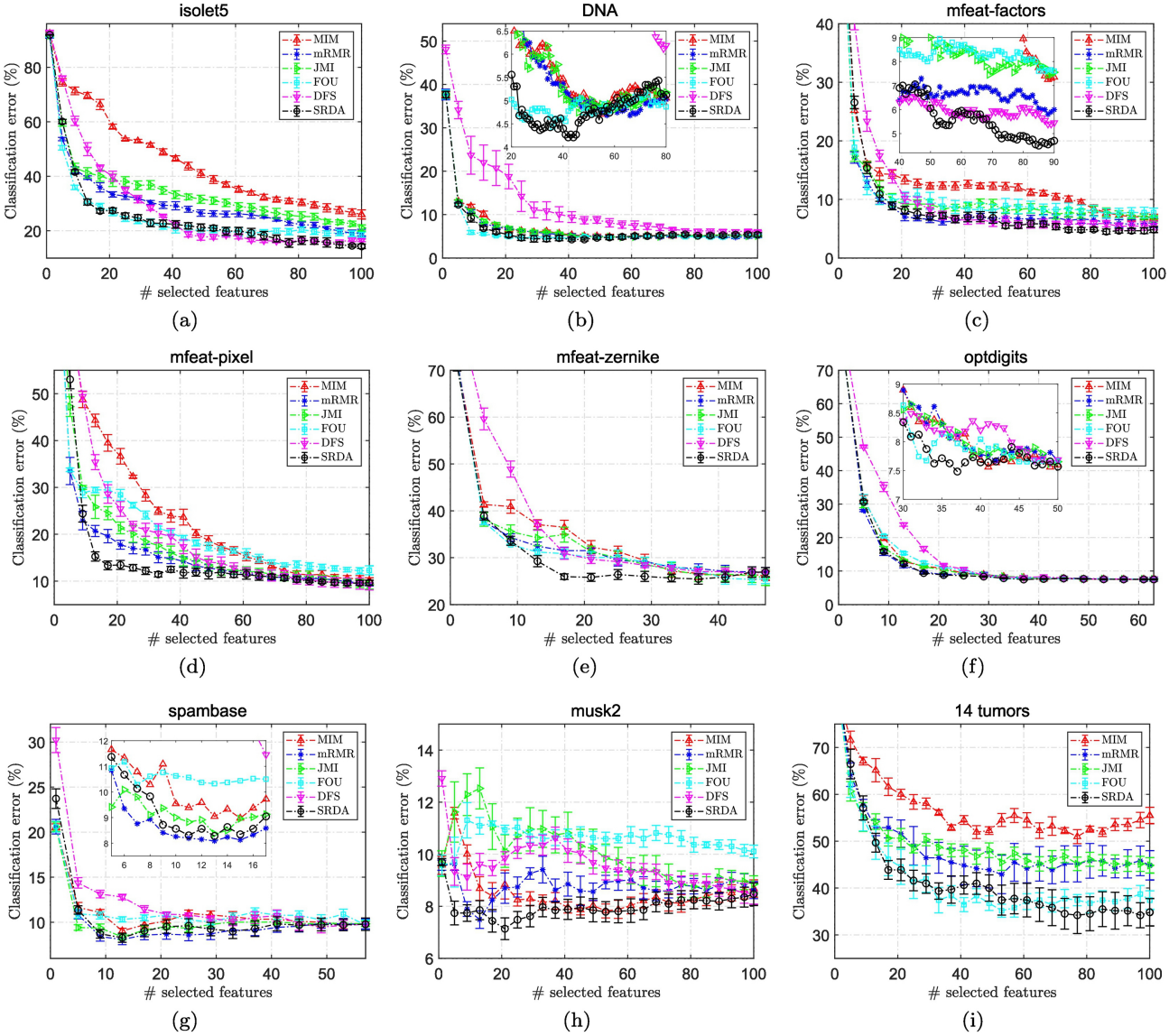
isolet5
DNA
mfeat-factors
mfeat-pixel
mfeat-zernike
optdigits
spambase
musk2
14 tumors
MIM
mRMR
JMI
FOU
DFS
SRDA
Classification error (%)
selected features
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)

Figure 5

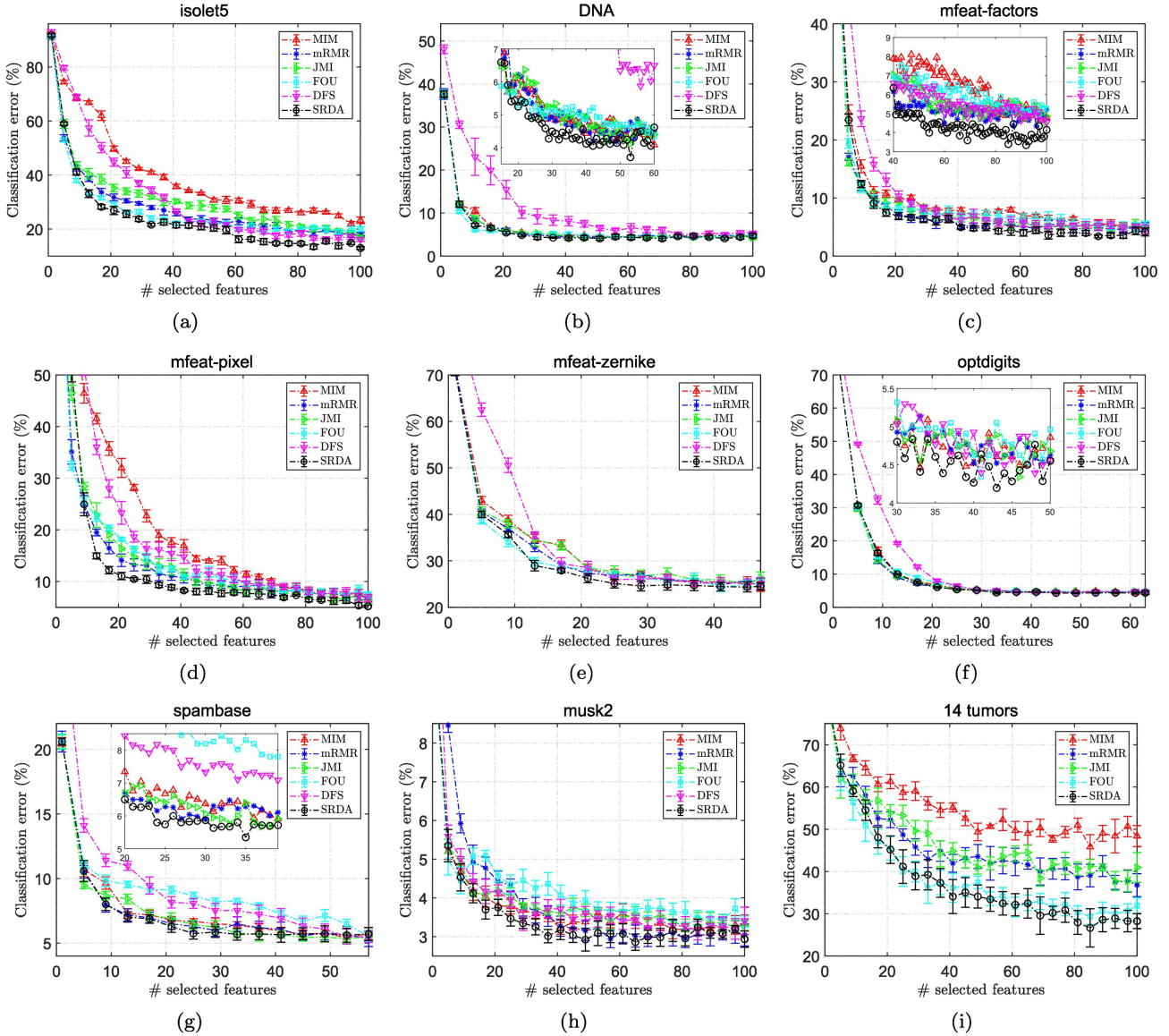
isolet5
DNA
mfeat-factors
mfeat-pixel
mfeat-zernike
optdigits
spambase
musk2
14 tumors
MIM
mRMR
JMI
FOU
DFS
SRDA
Classification error (%)
selected features
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)

Figure 6

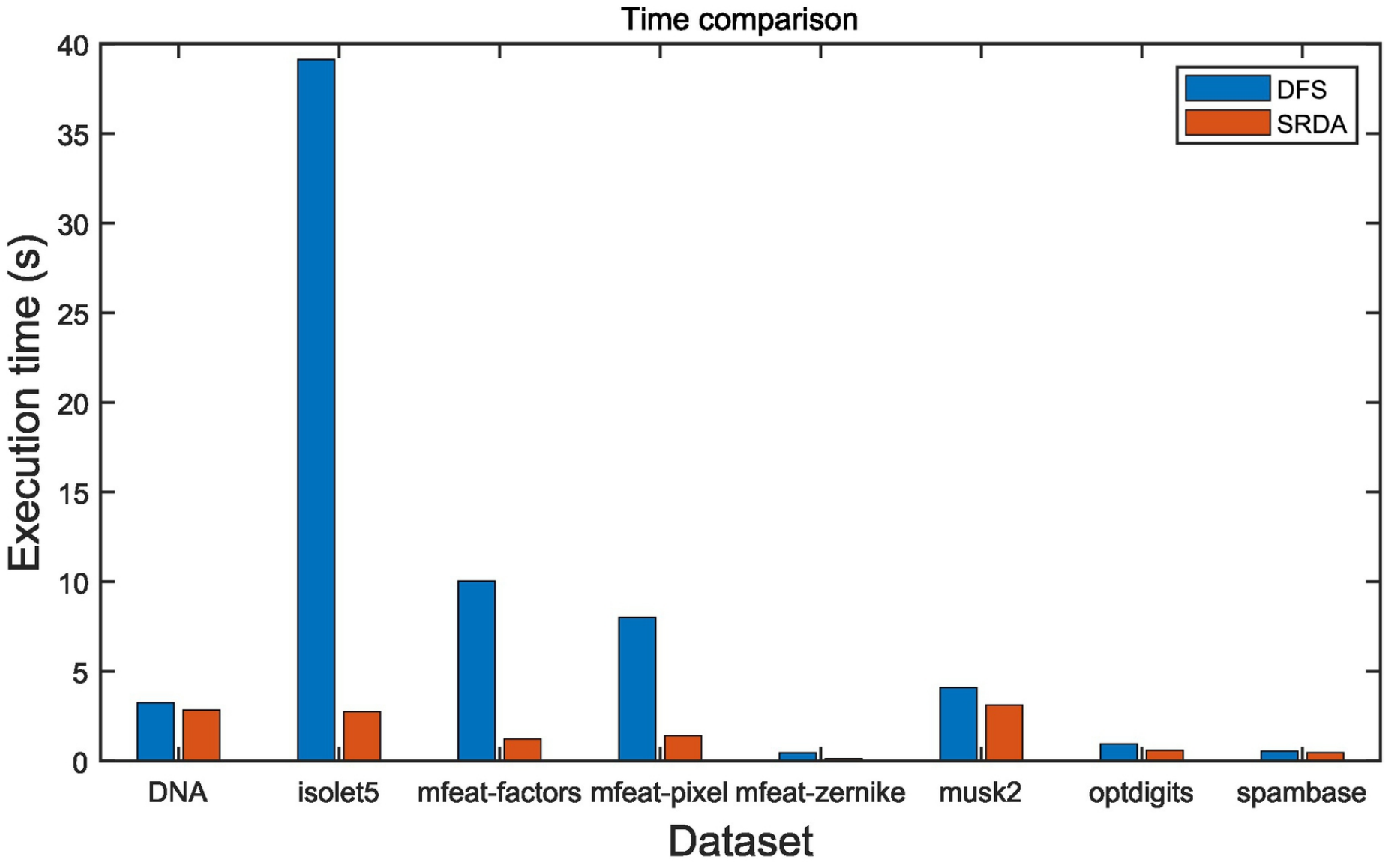
Time comparison
DFS
SRDA
Execution time (s)
0
5
10
15
20
25
30
35
40
DNA
isolet5
mfeat-factors
mfeat-pixel
mfeat-zernike
musk2
optdigits
spambase
Dataset

Figure 7