Annals of Operations Research (2021) 303:265–295
https://doi.org/10.1007/s10479-020-03917-w

S.I.: DATA MINING AND DECISION ANALYTICS

Solving a class of feature selection problems via fractional 0–1 programming

Erfan Mehmanchi[1] · Andrés Gómez[2] · Oleg A. Prokopyev[1]

Accepted: 22 December 2020 / Published online: 30 March 2021

Abstract
Feature selection is a fundamental preprocessing step for many machine learning and pattern recognition systems. Notably, some mutual-information-based and correlation-based feature selection problems can be formulated as fractional programs with a single ratio of polynomial 0–1 functions. In this paper, we study approaches that ensure globally optimal solutions for these feature selection problems. We conduct computational experiments with several real datasets and report encouraging results. The considered solution methods perform well for medium- and reasonably large-sized datasets, where the existing mixed-integer linear programs from the literature fail.

Keywords Feature selection · Fractional 0–1 programming · Mixed-integer linear programming · Parametric algorithms

1 Introduction

An essential preprocessing step for many data mining and machine learning tasks is the dataset dimensionality reduction that can be performed either by reducing the sizes of the sample or feature sets. In this paper, we focus on the latter procedure as a large number of features may cause model overfitting, which results in poor validation results (Chandrashekar and Sahin 2014; Jović et al. 2015).

Formally, a *feature* is a single measurable property of a process being observed. *Feature selection* is the process of identifying a subset of the most informative data features from the original feature set. Feature selection is often used in various machine learning and pattern

✉ Oleg A. Prokopyev
droleg@pitt.edu

Erfan Mehmanchi
erfan.mehmanchi@pitt.edu

Andrés Gómez
gomezand@usc.edu

[1] Department of Industrial Engineering, University of Pittsburgh, Pittsburgh, PA 15261, USA

[2] Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, CA 90089, USA

recognition settings that deal with large datasets including classification, clustering, and regression tasks. The corresponding applications arise in diverse areas such as e-commerce (Yuan et al. 2018), medical diagnosis (Fan and Chaovalitwongse 2010), bioinformatics (Saeys et al. 2007) and biomedicine (Busygin et al. 2008, 2005; Kocheturov et al. 2019), among others. Moreover, feature selection has other potential side benefits such as facilitating data visualization, decreasing training and utilization (computational) times as well as reducing the data storage requirements. We refer the readers to Guyon and Elisseeff (2003), Chandrashekar and Sahin (2014), Tang et al. (2014), Jović et al. (2015), and the references therein for an overview of applications and methods for feature selection.

In general, feature selection procedures are classified into three major categories, namely, filter, wrapper, and hybrid (embedded) methods (Chandrashekar and Sahin 2014; Jović et al. 2015). Wrapper and hybrid methods involve learning algorithms and the selection process is tailored based on the cho sen algorithm (El Ghaoui et al. 2010; Tibshirani et al. 2012; Viola et al. 2017; Atamtürk and Gómez 2020). In contrast, filter methods are not linked with any learning algorithm and are often a more appropriate choice for large-sized datasets (Nguyen et al. 2010b; Jović et al. 2015).

The main focus of this paper is on the filter methods. These methods select a subset of features by evaluating them according to some predefined measures. The measures typically applied in the literature can be categorized as information, distance, similarity, consistency, and statistical-based ones (Jović et al. 2015).

In this paper, we consider measures for the classification task in supervised learning, wherein we are given a training dataset and the classification of each sample is known. Then the goal is to predict unknown classes of new samples employing the information provided by the training dataset. To this end, it is important to distinguish *relevant* features from *redundant* ones, and hence, a desired measure (for feature selection) needs to differentiate the former from the latter. Relevant features are those that provide useful information for predicting the class of each given sample; redundant features are either weakly informative for this prediction or can be replaced with a set of some other relevant features.

The relevancy and redundancy are often characterized in terms of the correlation and mutual information concepts, which are widely used statistical tools (Peng et al. 2005). In particular, the studies in Hall (1999), Peng et al. (2005), and Ding and Peng (2005) propose a mutual-information-based and a correlation-based feature selection measures, referred to as *minimal redundancy maximal relevance* (mRMR) and *correlation feature selection* (CFS), respectively. A key advantage of these two approaches is that they take into account the features' relevancy and redundancy simultaneously.

Once a measure is selected, a procedure must be developed to select a subset of features from the full feature set. Finding an optimal subset, i.e., a subset that has the best value with respect to the considered measure (among exponentially many possible feature subsets) is often an NP-hard problem (Chandrashekar and Sahin 2014). Hence, in order to find a high quality (but not necessarily an optimal) subset, various heuristic methods have been proposed in the literature based on the mRMR and CFS measures; see, e.g., Yu and Liu (2003), Ding and Peng (2005), Peng et al. (2005), Brown et al. (2012), Huang et al. (2012), Liu and Motoda (2012), and Cilia et al. (2019). These heuristics are typically based on a (greedy) ranking of individual features with respect to the selected measure and then choosing a subset from the highest-ranking ones (Chandrashekar and Sahin 2014).

Nguyen et al. (2009, 2010b) show that the mRMR and CFS feature selection problems can be posed as single-ratio polynomial fractional 0–1 programs (PFPs), where the objective function is a ratio of quadratic binary functions. The existing exact solution approaches for the mRMR and CFS problems are centered around their transformations into equivalent *mixed-*

integer linear programs (MILPs). Notably, the PFPs of mRMR and CFS can be reformulated as MILPs by exploiting the methods from Chang (2001) and Nguyen et al. (2009); the latter method is also studied in Nguyen et al. (2010a, b, 2011). These reformulations are based on the substitution of the denominator of the ratio with a continuous variable and then linearizing the resulting quadratic and cubic terms, which, in turn, involve products of binary and at most one continuous variables. We also refer the reader to Mehmanchi et al. (2019) and Mehmanchi (2020) for additional details on reformulation approaches for fractional 0–1 programs.

On the other hand, the single-ratio structure of the PFP models of mRMR and CFS may allow us to use specialized approaches from fractional optimization instead of the generic MILP reformulations. In particular, an alternative approach can be based on parametric algorithms; see Borrero et al. (2017) and Ibaraki (1983) for reviews of such algorithms. Applying parametric algorithms to solve mRMR and CFS involves solving a sequence of binary quadratic problems (BQPs), which are also, in general, NP-hard (Palubeckis 2004). However, due to recent advances in binary quadratic optimization solvers that are readily available in CPLEX IBM (2019) and Gurobi Gurobi (2018), reasonably sized BQPs can be solved rather effectively. Additionally, within the parametric algorithms solving BQPs to optimality may not be required and each iteration of the algorithms can be stopped when a feasible solution satisfying some predefined conditions is found. This approach can lead to substantial improvements in the performance of these algorithms.

Contributions and the structure of the paper. The aim of this paper is to study exact approaches for the mRMR and CFS feature selection problems. To this end:

- In Sect. 2, we formally describe mRMR and CFS measures and the corresponding single-ratio fractional 0–1 optimization problems.
- In Sect. 3, first, we perform a comprehensive review of the existing MILP reformulations of the mRMR and CFS problems in the literature. Then by exploiting the structure of the fractional model of mRMR we propose a new MILP reformulation approach that outperforms the previous MILPs in the literature.
- In Sect. 4, we describe parametric methods from fractional optimization such as binary-search (Ahuja et al. 1993; Lawler 2001; Radzik 2013) and Newton's method (Dinkelbach 1967) algorithms that can be used for solving the mRMR and CFS problems.
- In Sect. 5, we conduct computational experiments with a collection of real datasets. From our results we observe that the performance of the existing MILPs in the literature is rather poor even for small- and medium-size problems. This observation is consistent with the earlier results in the literature (Nguyen et al. 2009, 2010b). On the other hand, the parametric methods perform well across all considered problem sizes. We also provide some insights on the selection of an appropriate measure and solution method.

2 Problem formulations

In the supervised learning for the purpose of classification the input data is given as an $n \times (p + 1)$ observation matrix, where n is the number of samples (observations). Each sample is a $(p + 1)$-dimensional vector of p features, $f_j,\ j \in J = \{1, 2, \ldots, p\}$, and the label of the class that contains this sample.

The aim of classification is to predict the label of the target class variable, denoted by C, for a given sample that indicates the classification of the sample. Then the feature selection problem is to find a subset $S \subseteq \{f_1, f_2, \ldots, f_p\}$ such that the reduced $n \times (|S| + 1)$ observation matrix provides sufficient information for the classification procedure to predict C.

Throughout the paper we let $\overline{C}$ denote the set of all possible labels for C, i.e., $C \in \overline{C}$. Next, we describe the mRMR and CFS feature selection measures and the corresponding optimization problems in Sects. 2.1 and 2.2, respectively.

2.1 mRMR optimization problem

In the information theory, the mutual information (MI) quantifies the amount of information that a random variable provides about another one and it can be used as a measure of the mutual dependency between two random variables (Peng et al. 2005). The notion of mutual information is related to the concept of entropy as the latter represents the "uncertainty" in the random variable. We refer to MacKay (2003) for more formal discussion on the entropy and mutual information.

Formally, let X and Y be two discrete random variables. Then the entropy of X is given as

$$\mathcal{H}(X) = -\sum_{x} \mathbb{P}(x) \log \mathbb{P}(x),$$

where $\mathbb{P}(x)$ is the probability that $X = x$. Moreover, the conditional entropy of X is given by

$$\mathcal{H}(X|Y) = -\sum_{x}\sum_{y} \mathbb{P}(x, y) \log \mathbb{P}(x|y),$$

which indicates the uncertainty that remains about X when we know the value of Y. Then the mutual information between X and Y, denoted by $\mathcal{I}(X, Y)$, is computed by

$$\mathcal{I}(X, Y) = \mathcal{H}(X) - \mathcal{H}(X|Y) = \mathcal{H}(Y) - \mathcal{H}(Y|X) = \sum_{x}\sum_{y} \mathbb{P}(x, y) \log \left[\frac{\mathbb{P}(x, y)}{\mathbb{P}(x)\mathbb{P}(y)}\right].$$

Note that $\mathcal{I}(X, Y)$ is non-negative; if X and Y are independent then $\mathcal{I}(X, Y)$ is zero and a larger value of $\mathcal{I}(X, Y)$ indicates larger dependency between X and Y. Additionally, observe that $\mathcal{I}(X, X) = \mathcal{H}(X)$. If X and Y are continuous variables, then similar definitions can be provided for $\mathcal{H}(X)$ and $\mathcal{I}(X, Y)$ by replacing the summations with integrations.

The task of feature selection using mRMR, proposed in Peng et al. (2005), is to find the subset $S \subseteq \{1, \ldots, n\}$ that has the maximum value for

$$\frac{1}{|S|}\sum_{f_j \in S} \mathcal{I}(f_j, C) - \frac{1}{|S|^2}\sum_{f_j, f_k \in S} \mathcal{I}(f_j, f_k), \tag{1}$$

over all 2^p possible feature subsets. The first term in (1) denotes the average MI between the features in set S and target class variable C, and thus, indicates the average relevancy of features in S. The second term denotes the average MI between features in S that also reflects the average redundancy of features in S.

In light of the above discussion, the maximization problem of (1) can be formulated as the fractional 0–1 program of the form, see Nguyen et al. (2009):

$$\text{(mRMR)} \qquad \max_{x \in \mathbb{B}^p} \left\{ \frac{\sum_{j\in J}\sum_{k\in J}\big(\mathcal{I}(f_j, C) - \mathcal{I}(f_j, f_k)\big)x_k x_j}{\sum_{j\in J}\sum_{k\in J} x_k x_j} \right\}, \tag{2}$$

where $\mathbb{B} := \{0, 1\}$ and $x_j = 1$ ($x_j = 0$) indicates the presence (absence) of feature f_j in set S.

2.2 CFS optimization problem

The mutual information is biased in favor of features that can take more number of values (Yu and Liu 2003). Moreover, for the purpose of comparing the degree of relevancy and redundancy of features, normalized values (i.e., adjusted values that have the same scale) are preferred. An alternative measure that can be used as an indicator of the relevancy and redundancy is correlation. In fact, a feature is said to be relevant if it is highly correlated with the target class variable, and it is redundant if it is highly correlated with some other features. These interpretations lead to the hypothesis that "good feature sets contain features that are highly correlated with the class, yet uncorrelated with each other" (Hall 1999).

The correlation—that is also referred to as symmetrical uncertainty (Yu and Liu 2003)—between two random variables X and Y can be obtained by their scaled MI (Press et al. 1992):

$$\mathcal{SU}(X,Y) = \frac{2\mathcal{I}(X,Y)}{\mathcal{H}(X)+\mathcal{H}(Y)},$$

which, in a sense, compensates the bias in MI. Also, $\mathcal{SU}(X,Y) \in [0,1]$, where 0 indicates the independency between X and Y and a larger value implies some degree of dependency.

Then feature selection by means of CFS, proposed in Hall (1999), is to find subset S which has the maximum value for:

$$\frac{\sum_{f_j\in S}\mathcal{SU}(f_j,C)}{\sqrt{|S| + 2\sum_{\substack{f_j,f_k\in S,\\ j\neq k}}\mathcal{SU}(f_j,f_k)}}. \tag{3}$$

Relation (3) provides the correlation of subset S and the target class. The numerator of (3) reflects the relevancy (correlation) of features in S to the target class; its denominator encompasses both the size of $|S|$ and the redundancy (inter-correlation) of features in S.

In view of the above discussion, the maximization problem of (3) over all 2^p possible feature subsets can be posed as the fractional 0–1 program of the form, see Nguyen et al. (2010a):

$$\text{(CFS)} \qquad \max_{x\in\mathbb{B}^p}\left\{\frac{\sum_{j\in J}\sum_{k\in J}\mathcal{SU}(f_j,C)\mathcal{SU}(f_k,C)x_kx_j}{\sum_{j\in J}x_j + \sum_{j\neq k}2\mathcal{SU}(f_j,f_k)x_kx_j}\right\}, \tag{4}$$

where $x_j = 1$ ($x_j = 0$) indicates the presence (absence) of feature f_j in set S.

3 Mixed-integer linear programming approaches

Both the mRMR and CFS feature selection problems given in (2) and (4), respectively, can be represented in the form of a single-ratio PFP as follows:

$$\lambda^\star = \max_{x\in\mathbb{B}^p}\frac{f(x)}{g(x)} := \max_{x\in\mathbb{B}^p}\left\{\frac{\sum_{j\in J}a_jx_j + \sum_{j\in J}\sum_{k\in J}b_{jk}x_jx_k}{\sum_{j\in J}c_jx_j + \sum_{j\in J}\sum_{k\in J}d_{jk}x_jx_k}\right\}, \tag{5}$$

where $a_j, b_{jk}, c_j, d_{jk} \in \mathbb{R}$, for all $j,k \in J := \{1,\ldots,p\}$. Moreover, the denominators of (2) and (4) are strictly positive whenever $|S| \geq 1$; the latter can enforced, if needed, by $\sum_{j\in J}x_j \geq 1$. Thus, throughout the paper we assume that $g(x) > 0$.

Herein, we first review the existing MILP solution methods to solve (5). In particular, first, we discuss the method proposed by Chang (2001) to transform (5) into MILP, which we denote as MILP$_1$; see Sect. 3.1. Second, we describe the approach of Nguyen et al. (2009), denoted by MILP$_2$; see Sect. 3.2. Finally, we propose a new MILP reformulations for the mRMR problem given in (2), which we denote by MILP$_3$; see Sect. 3.3.

3.1 Reformulation 1 (MILP$_1$)

We describe the approach of Chang (2001) for transforming PFPs into MILPs. To this end, define

$$y := \frac{1}{\sum_{j\in J} c_j x_j + \sum_{j\in J}\sum_{k\in J} d_{jk} x_j x_k}. \tag{6}$$

Then the substitution with variable y in (5) yields

$$\max_{x\in\mathbb{B}^p, y} \quad \sum_{j\in J} a_j x_j y + \sum_{j\in J}\sum_{k\in J} b_{jk} x_j x_k y \tag{7a}$$

$$\text{s.t.} \quad \sum_{j\in J} c_j x_j y + \sum_{j\in J}\sum_{k\in J} d_{jk} x_j x_k y = 1. \tag{7b}$$

Since $x_j \in \mathbb{B}$ and $x_k \in \mathbb{B}$, cubic terms $x_j x_k y$, for all $j, k \in J$, can be linearized as follows:

$$\Omega_{jk} := \left\{(x_j, x_k, y, z_{jk}) \in \mathbb{B}^2 \times \mathbb{R}^2 \mid \; 0 \le z_{jk} \le y^u x_j,\; z_{jk} \le y^u x_k,\; y^u(x_j + x_k - 2) + y \le z_{jk} \le y\right\},$$

where y^u is an upper bound on y; recall also that $y > 0$ by our assumption for the denominator of (5). Note that $(x_j, x_k, y, z_{jk}) \in \Omega_{jk} \Leftrightarrow z_{jk} = x_j x_k y$. Similarly, we use $\overline{\Omega}_j$ as a variant of Ω_{jk} to linearize bilinear (quadratic) terms $x_j y$, for all $j \in J$; specifically,

$$\overline{\Omega}_j := \left\{(x_j, y, \bar{z}_j) \in \mathbb{B} \times \mathbb{R}^2 \mid 0 \le \bar{z}_j \le y^u x_j,\; y^u(x_j - 1) + y \le \bar{z}_j \le y\right\}, \tag{8}$$

and $(x_j, y, \bar{z}_j) \in \overline{\Omega}_j \Leftrightarrow \bar{z}_j = x_j y$.

Hence, non-linear (due to the presence of terms $x_j x_k y$ and $x_j y$) and non-convex (for $x \in [0, 1]^p$) problem (7) is equivalent to the following MILP:

$$\begin{aligned}
(\text{MILP}_1) \qquad & \max && \sum_{j\in J} a_j \bar{z}_j + \sum_{j\in J}\sum_{k\in J} b_{jk} z_{jk} && \\
& \text{s.t.} && \sum_{j\in J} c_j \bar{z}_j + \sum_{j\in J}\sum_{k\in J} d_{jk} z_{jk} = 1 && \\
& && (x_j, x_k, y, z_{jk}) \in \Omega_{jk} && \forall j \le k \in J \\
& && (x_j, y, \bar{z}_j) \in \overline{\Omega}_j && \forall j \in J.
\end{aligned}$$

Let $a_j = c_j = 0,\; b_{jk} = \mathcal{I}(f_j, C) - \mathcal{I}(f_j, f_k)$, and $d_{jk} = 1$, for all $j, k \in J$, in MILP$_1$. Then we obtain an equivalent MILP of the mRMR feature selection problem (2). Similarly, in MILP$_1$, let $a_j = 0,\; b_{jk} = \mathcal{SU}(f_j, C) \cdot \mathcal{SU}(f_k, C)$, and $c_j = 1$, for all $j, k \in J$; additionally, set $d_{jk} = 2\mathcal{SU}(f_j, f_k)$, for $j \ne k \in J$ and $d_{jk} = 0$, for $j = k \in J$. Then we obtain an equivalent MILP of the CFS feature selection problem (4).

3.2 Reformulation 2 (MILP$_2$)

Nguyen et al. (2009) describe an alternative approach for transforming (5) into an MILP given as follows. Note that problem (7) can be rewritten as

$$\max_{x\in\mathbb{B}^p, y} \quad \sum_{j\in J} a_j x_j y + \sum_{j\in J} \big[(\sum_{k\in J} b_{jk} x_k) y\big] x_j \tag{9a}$$

$$\text{s.t.} \quad \sum_{j\in J} c_j x_j y + \sum_{j\in J}\big[(\sum_{k\in J} d_{jk} x_k) y\big] x_j = 1, \tag{9b}$$

where y is given in (6). Then define $v_j^b := [\sum_{k\in J} b_{jk} x_k y] x_j$ and $v_j^d := [\sum_{k\in J} b_{jk} x_k y] x_j$, for all $j \in J$. Observe that v_j^b and v_j^d are products of continuous terms, i.e., $\sum_{k\in J} b_{jk} x_k y$ and $\sum_{k\in J} d_{jk} x_k y$, respectively, and binary variable x_j.

Hence, in contrast to the approach of Sect. 3.1 that directly linearizes cubic terms $x_k x_j y$ using Ω_{ij}, by employing the technique used in $\overline{\Omega}_j$ we first replace cubic terms with a set of constraints involving linear and bilinear terms.

$$\max_{x\in\mathbb{B}^p, y, v, \bar{v}} \sum_{j\in J} a_j x_j y + \sum_{j\in J} v_j^b \tag{10a}$$

$$\text{s.t.} \quad \sum_{j\in J} c_j x_j y + \sum_{j\in J} v_j^d = 1 \tag{10b}$$

$$-\mathcal{M}_j^b x_j \leq v_j^b \leq \mathcal{M}_j^b x_j \qquad \forall j \in J \tag{10c}$$

$$\mathcal{M}_j^b (x_j - 1) + \sum_{k\in J} b_{jk} x_k y \leq v_j^b \leq \mathcal{M}_j^b (1 - x_j) + \sum_{k\in J} b_{jk} x_k y \qquad \forall j \in J \tag{10d}$$

$$-\mathcal{M}_j^d x_j \leq v_j^d \leq \mathcal{M}_j^d x \qquad \forall j \in J \tag{10e}$$

$$\mathcal{M}_j^d (x_j - 1) + \sum_{k\in J} d_{ij} x_k y \leq v_j^d \leq \mathcal{M}_j^d (1 - x_j) + \sum_{k\in J} d_{ij} x_k y \qquad \forall j \in J, \tag{10f}$$

where $\mathcal{M}_j^b$ and $\mathcal{M}_j^d$ are sufficiently large values for all $j \in J$. Then to transform (10) into an MILP we can linearize bilinear terms $x_k y$, for all $k \in J$ by using $\overline{\Omega}_j$. Thus, we get

$$
\begin{aligned}
(\text{MILP}_2) \max \quad & \sum_{j\in J} a_j \bar{z}_j + \sum_{j\in J} v_j^b \\
\text{s.t.} \quad & \sum_{j\in J} c_j \bar{z}_j + \sum_{j\in J} v_j^d = 1 \\
& \mathcal{M}_j^b (x_j - 1) + \sum_{k\in J} b_{jk} \bar{z}_k \leq v_j^b \leq \mathcal{M}_j^b (1 - x_j) + \sum_{k\in J} b_{jk} \bar{z}_k && \forall j \in J \\
& -\mathcal{M}_j^b x_j \leq v_j^b \leq \mathcal{M}_j^b x && \forall j \in J \\
& \mathcal{M}_j^d (x_j - 1) + \sum_{k\in J} d_{jk} \bar{z}_k \leq v_j^d \leq \mathcal{M}_j^d (1 - x_j) + \sum_{k\in J} d_{jk} \bar{z}_k && \forall j \in J \\
& -\mathcal{M}^d x_j \leq v_j^d \leq \mathcal{M}_j^d x_j && \forall j \in J \\
& (x_j, y, \bar{z}_j) \in \overline{\Omega}_j && \forall j \in J.
\end{aligned}
$$

Finally, to obtain equivalent MILPs of the mRMR and CFS feature selection problems (2) and (4), we need to use the same parameters settings, as described at the end of Sect. 3.1, in MILP_2.

3.3 New reformulation for mRMR (MILP_3)

Here, we propose a new MILP reformulation for the mRMR problem given in (2) based on its special structure. Notably, the denominator of the objective function ratio in problem (2), i.e., $\sum_{j\in J}\sum_{k\in J} x_j x_k$, takes values in the set $\{1^2, 2^2, 3^2 \dots, p^2\}$. Thus, using the standard value-disjunction approach we have:

$$\frac{1}{\sum_j \sum_k x_k x_j} = \sum_{\ell\in J} \frac{1}{\ell^2} w_\ell,$$

where $w_\ell \in \mathbb{B}$ with $\sum_{\ell\in J} w_\ell = 1$ and $\sum_{j\in J} x_j = \sum_\ell \ell w_\ell$. Thus, problem (2) reduces to:

$$\max_{x,w\in\mathbb{B}^p} \quad \sum_{\ell\in J}\sum_{j\in J}\sum_{k\in J} \frac{\mathcal{I}(f_j, C) - \mathcal{I}(f_j, f_k)}{\ell^2} x_k x_j w_\ell \tag{11a}$$

$$\text{s.t.} \quad \sum_{j\in J} x_j = \sum_{\ell\in J} \ell w_\ell \tag{11b}$$

$$\sum_{\ell\in J} w_\ell = 1. \tag{11c}$$

In order to transform (11) into an MILP, a possible approach is to define $u_{\ell jk} = x_k x_j w_\ell$ and use the technique of Glover and Woolsey (1974) to linearize cubic binary term $x_k x_j w_\ell$. The resulting MILP is:

$$\max_{x,w\in\mathbb{B}^p,u\geq 0} \quad \sum_{\ell\in J}\sum_{j\in J}\sum_{k\in J} \frac{\mathcal{I}(f_j, C) - \mathcal{I}(f_j, f_k)}{\ell^2} u_{\ell jk} \tag{12a}$$

$$\text{s.t.} \quad \sum_{j\in J} x_j = \sum_{\ell\in J} \ell w_\ell \tag{12b}$$

$$\sum_{\ell\in J} w_\ell = 1 \tag{12c}$$

$$u_{\ell jk} \leq w_\ell, u_{\ell jk} \leq x_j, u_{\ell jk} \leq x_k \qquad \forall \ell \in J, \forall j \leq k \in J \tag{12d}$$

$$u_{\ell jk} \geq w_\ell + x_j + x_k - 2 \qquad \forall \ell \in J, \forall j \leq k \in J. \tag{12e}$$

Based on our experiments, formulation (12) performs poorly in computations and does not scale well to large datasets. Nonetheless, as described next, an alternative linearization of the cubic terms leads to significantly better results.

In particular, we first transform the cubic expressions into bilinear terms, and then linearize the latter. Letting

$$r := \sum_{j\in J}\sum_{k\in J} \big(\mathcal{I}(f_j, C) - \mathcal{I}(f_j, f_k)\big) x_k x_j,$$

problem (11) reduces to:

$$\max_{x,w\in\mathbb{B}^p,r} \quad \sum_{\ell\in J} \frac{1}{\ell^2} r w_\ell \tag{13a}$$

$$\text{s.t.} \quad r = \sum_{j\in J}\sum_{k\in J} \big(\mathcal{I}(f_j, C) - \mathcal{I}(f_j, f_k)\big) x_k x_j \tag{13b}$$

$$\sum_{j\in J} x_j = \sum_{\ell\in J} \ell w_\ell \tag{13c}$$

$$\sum_{\ell \in J} w_\ell = 1. \tag{13d}$$

Next, we introduce continuous variable $t_{jk} := x_k x_j$ and use the technique of Glover and Woolsey (1974) to linearize binary quadratic term $x_k x_j$. Additionally, we define continuous variable $s_\ell := r w_\ell$ and use a variant of $\overline{\Omega}_j$ to linearize $r w_\ell$. Hence, we obtain:

$$\begin{aligned}
(\mathrm{MILP}_3) \qquad & \max_{x,w \in \mathbb{B}^p, t \geq 0, s, r} && \sum_{\ell \in J} \frac{1}{\ell^2} s_\ell \\
& \text{s.t.} && r = \sum_{j \in J} \sum_{k \in J} \big(\mathcal{I}(f_j, C) - \mathcal{I}(f_j, f_k)\big) t_{jk} \\
& && \sum_{j \in J} x_j = \sum_{\ell \in J} \ell w_\ell \\
& && \sum_{\ell \in J} w_\ell = 1 \\
& && t_{jk} \leq x_j,\ t_{jk} \leq x_k,\ t_{jk} \geq x_j + x_k - 1 \qquad \forall j \leq k \in J \\
& && s_\ell \leq \mathcal{M} w_\ell, s_\ell \leq r + \mathcal{M}(1 - w_\ell) \qquad \forall \ell \in J,
\end{aligned}$$

where $\mathcal{M}$ is a sufficiently large value. Note that since the MILP$_3$ is a maximization problem, the lower bounds for s_ℓ can be dropped.

4 Parametric approaches

Parametric algorithms are typical solution methods for single-ratio fractional (either binary or continuous) programs; see, e.g., reviews in Borrero et al. (2017) and Ibaraki (1983). Simply speaking, parametric algorithms find an optimal solution of a single-ratio fractional problem, as in (5), by solving a sequence of non-fractional problems.

Specifically, define parameter $t \in \mathbb{R}$ and consider the parametric optimization problem:

$$v(t) = \max_{x \in \mathbb{B}^p} \Big\{ f(x) - t \cdot g(x) \Big\}, \tag{15}$$

where $f(x)$ and $g(x)$ are defined as in (5). Note that for fixed t, problem (15) is a BQP as both $f(x)$ and $g(x)$ contain quadratic terms of binary variables.

Next, we observe that, under the positive denominator assumption, i.e., $g(x) > 0$, function $v(t)$ is monotone and if $v(t) = 0$, then t is the optimal objective function value of (5), i.e., $t = \lambda^\star$. Otherwise, we have either $v(t) > 0$ or $v(t) < 0$, which indicates, respectively, that $t < \lambda^\star$ and $t > \lambda^\star$. Thus, problem (5) reduces to the problem of finding a root of function $v(t)$.

Consequently, for our problems we can exploit root-finding methods that solve a sequence of either BQPs, or their equivalent linearized MILP versions. To obtain the latter, we observe that for (15) we can simply apply a variant of (8) to linearize the nonlinear terms $x_i x_j$; see, e.g., Glover and Woolsey (1974). Next, we first outline the binary-search method (Lawler 2001; Radzik 2013) in Sect. 4.1; then we describe the Newton-like method (Dinkelbach 1967; Megiddo 1979; Borrero et al. 2017) in Sect. 4.2.

4.1 Binary-search algorithm

Suppose that for the optimal objective function value $\lambda^\star$ at the beginning of iteration i of the algorithm an upper-bound, $\overline{\lambda}^i$, and a lower-bound, $\underline{\lambda}^i$, are given, i.e., it is known that $\lambda^\star \in [\underline{\lambda}^i, \overline{\lambda}^i]$. Then the binary-search algorithm (Lawler 2001; Radzik 2013) evaluates $v(\lambda^i_M)$, where λ^i_M is the midpoint of the given interval, i.e., $\lambda^i_M = (\underline{\lambda}^i + \overline{\lambda}^i)/2$. If $v(\lambda^i_M) > 0$, then we update the lower-bound, $\underline{\lambda}^{i+1} = \lambda^i_M$; if $v(\lambda^i_M) < 0$, then we update upper-bound, $\overline{\lambda}^{i+1} = \lambda^i_M$; else, we have $v(\lambda^i_M) = 0$ and the midpoint λ^i_M is the optimal objective function value. The formal pseudo-code is given in Algorithm 1 below.

Algorithm 1 Binary-search algorithm

```
1: Input: $\epsilon_{rel}$, relative gap parameter; $\epsilon_{abs}$, absolute gap parameter
2: Output: $x$; if $x_j = 1$, then feature $j$ is selected
3: $i \leftarrow 0$
4: Compute $\overline{\lambda}^0$ and $\underline{\lambda}^0$
5: while time limit not exceeded & $|(\overline{\lambda}^i - \underline{\lambda}^i)/\underline{\lambda}^i| > \epsilon_{rel}$ & $|\overline{\lambda}^i - \underline{\lambda}^i| > \epsilon_{abs}$ do
6:     $\lambda^i_M \leftarrow (\underline{\lambda}^i + \overline{\lambda}^i)/2$
7:     Solve problem (15) for $t = \lambda^i_M$ and obtain $v(\lambda^i_M)$ and its optimal solution $x^i$
8:     if $v(\lambda^i_M) > 0$ then
9:         $\underline{\lambda}^{i+1} \leftarrow \lambda^i_M$, $\overline{\lambda}^{i+1} \leftarrow \overline{\lambda}^i$
10:    else if $v(\lambda^i_M) < 0$ then
11:        $\underline{\lambda}^{i+1} \leftarrow \underline{\lambda}^i$, $\overline{\lambda}^{i+1} \leftarrow \lambda^i_M$
12:    else
13:        return $x^i$                                    ▷ Optimal solution found
14:    end if
15:    $i \leftarrow i + 1$
16: end while
17: return $x^i$                              ▷ Best solution found within the time limit
```

Note that at each iteration of Algorithm 1 we do not need to solve problem (15) in line 7 to optimality; instead, we can stop whenever a feasible solution with a positive objective function value is found. This observation can potentially result in a better performance for the binary-search algorithm. In fact, mixed integer optimization algorithms often find feasible and even optimal solutions in a portion of the time required to prove the optimality. Thus, if problem (15) is solved until the first feasible solution with a positive objective function value is found, then in practice most of the iterations, except the few last, are solved with a few branch-and-bound nodes. Although this approach may require more iterations, the total solution times are often improved significantly.

We define $h(x) := \frac{f(x)}{g(x)}$. Thus,

$$\lambda^\star = \max_{x \in \mathbb{B}^p} h(x) = \max_{x \in \mathbb{B}^p} \frac{f(x)}{g(x)}. \tag{16}$$

Next, let $x^\star$ denote an optimal solution of (16), i.e., $x^\star \in \operatorname*{argmax}_{x \in \mathbb{B}^p} h(x)$. Then for any feasible solution $\bar{x}$ we define the relative and absolute optimality gaps as follows.

$$\text{Relative gap: } \mathtt{Gap}_{rel} := \left| \frac{h(x^\star) - h(\bar{x})}{h(\bar{x})} \right|, \qquad \text{Absolute gap: } \mathtt{Gap}_{abs} := |h(x^\star) - h(\bar{x})|. \tag{17}$$

If Algorithm 1 terminates before reaching the time limit, then it yields a feasible solution with either $\texttt{Gap}_{rel} \leq \epsilon_{rel}$ or $\texttt{Gap}_{abs} \leq \epsilon_{abs}$. If the time limit is reached after processing the i-th iteration of the algorithm, then

$$\texttt{Gap}_{rel} \leq |(\overline{\lambda}^i - \underline{\lambda}^i)/\underline{\lambda}^i|, \text{ and } \texttt{Gap}_{abs} \leq |\overline{\lambda}^i - \underline{\lambda}^i|. \tag{18}$$

4.2 Newton-like method algorithm

The second approach that we employ to find the root of problem (15) is based on Newton-like method (Dinkelbach 1967; Megiddo 1979; Borrero et al. 2017) described as follows. Suppose that at the beginning of iteration i a lower-bound t^i on $\lambda^\star$ is known, which can be obtained, e.g., by computing the fractional objective function at any feasible solution. If $v(t^i) = 0$, then $t^i = \lambda^\star$; otherwise, the algorithm updates $t^{i+1} = h(x^i)$, where x^i is an optimal solution of $v(t^i)$, and proceeds to the next iteration. The formal pseudo-code is given in Algorithm 2.

Algorithm 2 Newton-like method algorithm

```
1: Input: $\epsilon_{rel}$, relative gap parameter; $\epsilon_{abs}$, absolute gap parameter
2: Output: $x$; if $x_j = 1$, then feature $j$ is selected
3: $i \leftarrow 0$
4: Compute $t^i$                                        ▷ e.g., $t^i = h(\mathbf{1}')$
5: while time limit not exceeded do
6:     Solve problem (15) for $t^i$ and obtain $v(t^i)$ and its optimal solution $x^i$
7:     if $v(t^i) > \epsilon_{rel} \cdot |t^i|$ and $v(t^i) > \epsilon_{abs}$ then
8:         $t^{i+1} \leftarrow h(x^i)$
9:     else
10:        return $x^i$                ▷ Solution found within either relative or optimality gaps
11:    end if
12:    $i \leftarrow i + 1$
13: end while
14: return $x^i$                                ▷ Best solution found within the time limit
```

Note that at each iteration of Algorithm 2 we can stop the optimization of problem (15) in line 6 whenever a feasible solution with an objective function value greater than $\epsilon_{rel} \cdot |t^i|$ and ϵ_{abs} is found. Similar to the observation made in Sect. 4.1, this modification of the algorithm can result in more iterations but a better overall performance.

Recall the relative and optimality gaps defined in (17). Following the proofs of similar results in Gómez and Prokopyev (2020) - Proposition 4 - and Radzik (2013) if the time limit is not reached, then Algorithm 2 terminates with a feasible solution with either $\texttt{Gap}_{rel} \leq \epsilon_{rel}$ or $\texttt{Gap}_{abs} \leq \epsilon_{abs}$. If the time limit is reached after the operation of the i-th iteration of Algorithm 2, then we estimate relative and absolute gaps by

$$\texttt{Gap}_{rel} \simeq \frac{v(t^i)}{|t^i| \cdot g(x^i)}, \text{ and } \texttt{Gap}_{abs} \simeq \frac{v(t^i)}{g(x^i)}. \tag{19}$$

5 Computational results

The aim of our computational study is to evaluate the performances of the MILP reformulations provided in Sect. 3 versus the parametric approaches of Sect. 4. Note that results

Table 1 Considered datasets

Dataset	Abbreviated name	p	n	Data type	Class type
Banknote_authentication[a]	banknote_auth	4	1372	Continuous	Binary
Breast_cancer[a]	Breast_cancer	9	286	Discrete	Binary
Letter_Recognition[a]	Letter_Recog	16	20000	Discrete	Multi
Zoo[a]	Zoo	17	101	discrete	multi
Breast_Cancer_Wisconsin_(Diagnostic)[a]	Breast_Cancer	31	569	Continuous	Binary
SPECTF_Heart_Data[a]	SPECTF_Heart	44	267	Continuous	Binary
Lung_Cancer[a]	Lung_Cancer	56	32	Discrete	Binary
Sports_articles_for_objectivity_analysis[a]	Sports_articles	59	1000	Discrete	Binary
Connectionist[a]	Connectionist	60	208	Continuous	Binary
Optical_Recognition[a]	Optical_Recog	62	3823	Discrete	Multi
Hill-Valley[a]	Hill-Valley	100	606	Continuous	Binary
Urban_Land_Cover[a]	Urban_Land	147	168	Continuous	Multi
Epileptic_Seizure_Recognition[a]	Epileptic_Seiz	178	11500	Discrete	Multi
SCADI[a]	SCADI	205	70	Discrete	Multi
Semeion_Handwritten_Digit[a]	Semeion_Hand	256	1593	Discrete	Multi
USPS[b]	USPS	256	9298	Continuous	Multi
lung_discrete[b]	lung_discrete	325	73	Discrete	Multi
Madelon[1,2]	Madelon	500	2000	Continuous	Binary
ISOLET[1,2]	ISOLET	617	7797	Continuous	Multi
Parkinson's_Disease[a]	Parkinson	754	756	Continuous	Binary
CNAE-9[a]	CNAE-9	856	1080	Discrete	Multi
Yale_32x32[b]	Yale_32x32	1024	165	Continuous	Multi
ORL_32x32[b]	ORL_32x32	1024	400	Continuous	Multi
colon[b]	Colon	2000	62	Discrete	Binary
PCMAC[b]	PCMAC	3289	1943	Discrete	Binary

We provide the number of features, p, and the number of samples, n as well as the types of the features' values and the target class variable. For the latter, if $|\overline{C}| = 2$, then the target class is binary, otherwise, it is multi-class. Datasets' abbreviated names are used in Tables 2, 3, 4, 5, 6 and 7

[a] UCI machine learning repository (Asuncion and Newman 2007)

[b] ASU feature selection repository (Li et al. 2016)

for the MILP formulation (12) are omitted as it was consistently outperformed by the other approaches. In Sect. 5.1, we outline the real-life datasets and the parameter settings used for the computational experiments. Then we present our results in Sect. 5.2.

5.1 Computational environment and datasets

For all considered datasets, we solve MILPs and BQPs (at each iteration of the parametric Algorithms 1 and 2) using CPLEX 12.9.0 IBM (2019). We run experiments on a PC with 32-core CPU (2.90 GHz) and 160 GB of RAM; we allocate 4 threads and 16 GB of RAM for each individual experiment. We use the time limit of one hour (3600 seconds). To avoid running-out-of-memory difficulties we use the "node-file storage-feature" of CPLEX to store some parts of the branch-and-cut tree on a disk when the size of the tree exceeds the allocated memory. Furthermore, for computing the mutual information and correlation between a feature and the target class or between two features, as well as computing F1 score we use *scikit-learn* package (Pedregosa et al. 2011) and Python 3.7.7 (Python Software Foundation. 2020).

Datasets We consider various real-world datasets obtained from *UCI machine learning repository* (Asuncion and Newman 2007) and *ASU feature selection repository* (Li et al. 2016) available at https://archive.ics.uci.edu and http://featureselection.asu.edu, respectively. Table 1 provides the list of the datasets as well as their sizes and key characteristics.

Linearization bounds In both MILP$_1$ and MILP$_2$, we let $y^u = 1$. Moreover, for MILP$_2$ reformulation of mRMR we let $\mathcal{M}_j^b = \sum_{k \in J} |\mathcal{I}(f_j, C) - \mathcal{I}(f_j, f_k)|$ and $\mathcal{M}_j^d = n$, for all $j \in J$. For MILP$_2$ reformulation of CFS we set $\mathcal{M}_j^b = \sum_{k \in J} \mathcal{SU}(f_j, C) \cdot \mathcal{SU}(f_k, C)$ and $\mathcal{M}_j^d = \sum_{k \in J, k \neq j} 2\mathcal{SU}(f_j, f_k)$, for all $j \in J$. Finally, we set $\mathcal{M} = \sum_{j \in J} \sum_{k \in J} |\mathcal{I}(f_j, C) - \mathcal{I}(f_j, f_k)|$ in MILP$_3$.

Gaps We consider $\epsilon_{rel} = 0.01$ and $\epsilon_{abs} = 0.001$ in both Algorithms 1 and 2. If the time limit is reached, then $\texttt{Gap}_{rel}$ and $\texttt{Gap}_{abs}$ are computed by using formulas given in (18) and (19) for Algorithms 1 and 2, respectively. Similarly, we set 0.01 and 0.001 for the relative and absolute optimality gaps in the MIP solver which are computed by $\texttt{Gap}_{rel} = |\frac{UB-LB}{LB}|$ and $\texttt{Gap}_{abs} = |UB - LB|$, where UB and LB are the upper- and the lower-bounds on the optimal objective function value at the termination of the solver, respectively.

Classification score We evaluate a subset of features in predicting the true class of samples in the dataset by F1 score. To this end, we use the well-known *Naive Bayes* and *Random Forest* classifiers (commonly used in the related literature, see, e.g., Peng et al. (2005) and Nguyen et al. (2009, 2010a), with the 5-fold cross validation.

Heuristic methods In order to further evaluate the computational and classification performances of the considered exact solution methods, we perform computational experiments with heuristic methods for the mRMR and CFS feature selection problems based on the approaches in Brown et al. (2012) and Zhao et al. (2010), respectively. Specifically, we use the implementations from Li et al. (2016).

5.2 Results and analysis

Next, we evaluate the computational and classification performances of the MILPs in Sect. 3 versus Algorithms 1 and 2 in Sect. 4 as well as the considered heuristic methods from the literature. In particular, the computational results are presented in Tables 2, 3 and 4 and the classification results are reported in Tables 5, 6 and 7.

First, we discuss the computational results for the MILPs in solving the mRMR feature selection problem, see Table 2. We observe that for "small" datasets ($p \leq 60$), MILP$_3$ has, in general, the best performance among the MILPs. In particular, for most of the datasets with $44 \leq p \leq 60$, MILP$_1$ and MILP$_2$ do not find an optimal solution within the time limit, while MILP$_3$ solves the same datasets to optimality in only a few seconds. For larger datasets ($p > 60$), all MILPs reach the time limit. For these datasets, MILP$_1$ often struggles to find a feasible solution, while MILP$_2$ and MILP$_3$ report large gaps, see Table 2. Thus, we conclude that while MILP formulations—particularly MILP$_3$—are adequate at tackling problems with small number of features, they struggle with high-dimensional datasets.

Table 2 Computational performance of MILP$_1$, MILP$_2$ and MILP$_3$ when solving the mRMR feature selection problem (2)

Dataset	p	MILP$_1$ (Chang 2001)			MILP$_2$ (Nguyen et al. 2009)			MILP$_3$		
		Gap$_{abs}$	Gap$_{rel}$	Time	Gap$_{abs}$	Gap$_{rel}$	Time	Gap$_{abs}$	Gap$_{rel}$	Time
banknote_auth	4	0.000	0.00	0.1	0.000	0.00	**0.0**	0.000	0.00	**0.0**
Breast_cancer	9	0.000	0.00	**0.1**	0.000	0.00	**0.1**	0.000	0.00	**0.1**
Letter_Recog	16	0.002	0.01	1.3	0.002	0.01	2.0	0.001	0.00	**0.2**
Zoo	17	0.003	0.01	**0.7**	0.003	0.01	5.0	0.003	0.01	1.1
Breast_Cancer	31	0.001	0.03	176.7	0.001	0.03	20.2	0.001	0.03	**9.8**
SPECTF_Heart	44	0.024	0.34	T	1.607	22.85	T	0.001	0.01	**55.1**
Lung_Cancer	56	0.027	2.77	T	3.432	+	T	0.001	0.10	**17.8**
Sports_articles	59	0.001	+	53.8	0.083	+	T	0.000	0.00	**1.4**
Connectionist	60	0.001	0.79	T	0.004	2.14	T	0.001	0.55	**21.6**
Optical_Recog	62	**0.064**	0.34	T	5.822	30.91	T	0.291	1.54	T
Hill-Valley	100	**0.037**	0.06	T	0.597	0.95	T	0.630	1.00	T
Urban_Land	147	**0.273**	1.14	T	+	+	T	7.871	31.44	T
Epileptic_Seiz	178	**0.019**	0.36	T	+	+	T	2.426	44.97	T
SCADI	205	**0.096**	0.39	T	+	+	T	3.589	14.74	T
Semeion_Hand	256	**0.072**	0.49	T	+	+	T	+	+	T

Table 2 continued

Dataset	p	MILP$_1$ (Chang 2001)			MILP$_2$ (Nguyen et al. 2009)			MILP$_3$		
		Gap$_{abs}$	Gap$_{rel}$	Time	Gap$_{abs}$	Gap$_{rel}$	Time	Gap$_{abs}$	Gap$_{rel}$	Time
USPS	256	**0.137**	0.70	T	+	+	T	+	+	T
lung_discrete	325	**0.236**	1.19	T	+	+	T	+	+	T
Madelon	500	–	–	T	**0.127**	+	T	1.131	58.16	T
ISOLET	617	**0.114**	5.85	T	+	+	T	+	+	T
Parkinson	754	–	–	T	+	+	T	**9.500**	+	T
CNAE-9	856	–	–	T	21.770	+	T	**0.122**	+	T
Yale_32x32	1024	–	–	T	+	+	T	+	+	T
ORL_32x32	1024	–	–	T	+	+	T	+	+	T
colon	2000	–	–	T	+	+	T	+	+	T
PCMAC	3289	–	–	T	+	+	T	+	+	T

For each dataset and solution method we report the absolute (Gap$_{abs}$) and relative (Gap$_{rel}$) gaps, as well as the running time (Time, in seconds). For each dataset, among the solution methods, the best Time and the best Gap$_{abs}$ (if Time$\geq$ 3600 sec) are in **bold**, and p indicates the size of the full feature set
"–": No feasible solution is found within the time limit. "+": Gap is larger than 100. "T": Time limit (3600 sec.) is reached

Table 3 Computational performance of the best MILPs (MILP_1 and MILP_3), Algorithms 1 and 2, and the heuristic method (Heu.) when solving the mRMR feature selection problem (2)

Dataset	p	MILP_1 (Chang 2001)			MILP_3			Algorithm 1				Algorithm 2				Heu. (Brown et al. 2012)
		Gap_{abs}	Gap_{rel}	Time	Gap_{abs}	Gap_{rel}	Time	Gap_{abs}	Gap_{rel}	Time	#	Gap_{abs}	Gap_{rel}	Time	#	Time
banknote_auth	4	0.000	0.00	0.1	0.000	0.00	**0.0**	0.004	0.01	0.4	10	0.001	0.01	0.1	2	0.1
Breast_cancer	9	0.000	0.00	**0.1**	0.000	0.00	**0.1**	0.001	0.01	0.5	12	0.001	0.01	0.2	3	0.0
Letter_Recog	16	0.002	0.01	1.3	0.001	0.00	**0.2**	0.002	0.01	0.7	11	0.001	0.01	**0.2**	3	2.6
Zoo	17	0.003	0.01	0.7	0.003	0.01	1.1	0.003	0.01	0.5	10	0.001	0.01	**0.2**	5	0.1
Breast_Cancer	31	0.001	0.03	176.7	0.001	0.03	**9.9**	0.001	0.02	18.3	12	0.001	0.01	**9.9**	7	0.2
SPECTF_Heart	44	0.024	0.34	T	0.001	0.01	55.1	0.001	0.01	**0.8**	12	0.001	0.01	**0.8**	7	0.1
Lung_Cancer	56	0.027	2.77	T	0.001	0.10	17.8	0.001	0.08	**1.1**	12	0.001	0.01	1.3	8	0.1
Sports_articles	59	0.001	+	53.8	0.000	0.00	1.4	0.001	3.13	**0.3**	12	0.001	0.01	**0.3**	11	0.7
Connectionist	60	0.001	0.79	T	0.001	0.55	21.6	0.001	0.47	**12.5**	12	0.001	0.01	22.8	5	0.1
Optical_Recog	62	0.064	0.34	T	0.291	1.54	T	0.001	0.01	3.5	11	0.001	0.01	**2.6**	5	27.6
Hill-Valley	100	0.037	0.06	T	0.630	1.00	T	0.004	0.01	**3.4**	10	0.001	0.01	7.8	10	0.7
Urban_Land	147	0.273	1.14	T	7.871	31.44	T	0.087	0.42	T	5	**0.027**	0.11	T	4	0.3
Epileptic_Seiz	178	0.019	0.36	T	2.426	44.97	T	0.003	0.05	T	10	**0.003**	0.05	T	17	14.2
SCADI	205	0.096	0.39	T	3.589	14.74	T	0.001	0.01	**1186.2**	11	0.001	0.01	2524.5	16	6.7
Semeion_Hand	256	0.072	0.49	T	+	+	T	0.001	0.01	**145.9**	11	0.001	0.01	499.6	19	182.5
USPS	256	0.137	0.70	T	+	+	T	0.023	0.12	T	7	**0.020**	0.11	T	14	22.3
lung_discrete	325	0.236	1.19	T	+	+	T	0.003	0.01	**2785.9**	10	0.000	0.00	T	20	19.4
Madelon	500	–	–	T	1.131	58.16	T	0.001	0.04	**1683.7**	12	0.000	0.01	T	1	5.9

Table 3 continued

Dataset	p	MILP$_1$ (Chang 2001)			MILP$_3$			Algorithm 1				Algorithm 2				Heu. (Brown et al. 2012)
		$\mathtt{Gap}_{abs}$	$\mathtt{Gap}_{rel}$	`Time`	$\mathtt{Gap}_{abs}$	$\mathtt{Gap}_{rel}$	`Time`	$\mathtt{Gap}_{abs}$	$\mathtt{Gap}_{rel}$	`Time`	#	$\mathtt{Gap}_{abs}$	$\mathtt{Gap}_{rel}$	`Time`	#	`Time`
ISOLET	617	0.114	5.85	T	+	+	T	0.047	5.18	T	6	**0.003**	0.17	T	3	48.3
Parkinson	754	–	–	T	9.500	+	T	0.024	1.30	T	7	**0.001**	0.78	T	3	5.8
CNAE-9	856	–	–	T	0.122	+	T	0.012	1.17	T	8	**0.000**	0.00	T	3	62.9
Yale_32x32	1024	–	–	T	+	+	T	0.091	1.17	T	5	**0.031**	0.35	T	1	1.7
ORL_32x32	1024	–	–	T	+	+	T	0.092	1.28	T	5	**0.028**	0.35	T	1	3.7
colon	2000	–	–	T	+	+	T	0.097	0.86	T	5	**0.033**	0.32	T	1	38.0
PCMAC	3289	–	–	T	+	+	T	0.024	2.72	T	7	**0.000**	0.03	T	1	3521.5

For each dataset and solution method we report the absolute ($\mathtt{Gap}_{abs}$) and relative ($\mathtt{Gap}_{rel}$) gaps, as well as the running time (`Time`, in seconds). We also report the number of iterations for the algorithms, denoted by `#`. For each dataset, among the exact solution methods, the best `Time` and the best $\mathtt{Gap}_{abs}$ (if `Time`≥ 3600 sec) are in **bold**, and p indicates the size of the full feature set.

"–": No feasible solution is found within the time limit. "+": `Gap` is larger than 100. "T": Time limit (3600 sec.) is reached

Table 4 Computational performance of MILP$_1$ and MILP$_2$, Algorithms 1 and 2, and the heuristic method (Heu.) when solving the CFS feature selection problem (4)

Dataset	p	MILP$_1$ (Chang 2001)			MILP$_2$ (Nguyen et al. 2009)			Algorithm 1				Algorithm 2				Heu. (Zhao et al. 2010)
		Gap$_{abs}$	Gap$_{rel}$	Time	Gap$_{abs}$	Gap$_{rel}$	Time	Gap$_{abs}$	Gap$_{rel}$	Time	#	Gap$_{abs}$	Gap$_{rel}$	Time	#	Time
banknote_auth	4	0.000	0.00	**0.0**	0.000	0.00	**0.0**	0.001	0.01	0.3	12	0.001	0.01	0.1	3	0.4
Breast_cancer	9	0.000	0.02	**0.1**	0.001	0.04	**0.1**	0.001	0.04	0.3	12	0.001	0.01	**0.1**	3	0.2
Letter_Recog	16	0.001	0.06	0.3	0.001	0.06	**0.2**	0.001	0.05	0.4	12	0.001	0.01	**0.2**	4	142.4
Zoo	17	0.005	0.01	1.4	0.006	0.01	0.4	0.005	0.01	0.6	9	0.001	0.01	**0.3**	5	1.6
Breast_Cancer	31	0.000	0.00	**0.2**	0.001	0.01	**0.2**	0.001	0.01	0.4	11	0.001	0.01	0.3	7	14.2
SPECTF_Heart	44	0.006	0.01	274.4	0.293	0.53	T	0.005	0.01	**0.2**	9	0.001	0.01	0.6	8	75.6
Lung_Cancer	56	0.002	0.01	22.1	0.002	0.01	723.9	0.001	0.01	0.5	11	0.001	0.01	**0.3**	5	2.1
Sports_articles	59	0.001	0.01	3.1	0.001	0.01	1817.3	0.001	0.01	**0.2**	12	0.001	0.01	0.3	7	114.0
Connectionist	60	0.001	0.10	3.3	0.001	0.10	0.2	0.001	0.07	0.3	12	0.001	0.01	**0.1**	4	5.4
Optical_Recog	62	0.153	0.72	T	0.312	1.46	T	0.001	0.01	6.0	11	0.001	0.01	**2.0**	4	T
Hill-Valley	100	0.001	1.32	210.8	0.002	2.35	T	0.001	1.81	**0.3**	12	0.001	0.01	0.4	10	26.4
Urban_Land	147	1.001	3.77	T	2.546	9.59	T	0.022	0.09	T	7	**0.011**	0.05	T	6	16.3
Epileptic_Seiz	178	1.170	16.44	T	1.795	23.22	T	**0.001**	0.02	T	11	0.002	0.02	T	4	T
SCADI	205	7.526	16.38	T	2.569	5.55	T	0.011	0.03	T	8	**0.000**	0.00	T	9	85.4
Semeion_Hand	256	28.695	+	T	2.558	9.70	T	0.176	0.93	T	4	**0.061**	0.29	T	4	T
USPS	256	17.125	+	T	4.150	18.99	T	0.179	1.33	T	4	**0.025**	0.17	T	6	2764.3

Table 4 continued

Dataset	p	MILP$_1$ (Chang 2001)			MILP$_2$ (Nguyen et al. 2009)			Algorithm 1				Algorithm 2				Heu. (Zhao et al. 2010)
		Gap$_{abs}$	Gap$_{rel}$	Time	Gap$_{abs}$	Gap$_{rel}$	Time	Gap$_{abs}$	Gap$_{rel}$	Time	#	Gap$_{abs}$	Gap$_{rel}$	Time	#	Time
lung_discrete	325	14.884	57.82	T	14.878	27.75	T	0.335	1.04	T	3	**0.073**	0.22	T	4	3560.1
Madelon	500	0.007	11.04	T	0.013	9.52	T	0.001	0.57	**5.8**	12	0.001	0.01	7.1	8	512.7
ISOLET	617	+	+	T	1.154	16.87	T	0.024	0.01	**4.5**	7	0.344	3.21	T	1	T
Parkinson	754	+	+	T	0.541	19.80	T	0.024	0.01	**352.0**	7	0.134	1.53	T	4	201.5
CNAE-9	856	–	–	T	0.003	0.01	444.4	0.001	0.01	17.1	11	0.001	0.01	**14.2**	7	765.5
Yale_32x32	1024	–	–	T	49.784	+	T	0.090	0.84	T	5	**0.000**	0.00	T	2	1945.8
ORL_32x32	1024	–	–	T	29.502	+	T	0.046	0.60	T	6	**0.014**	0.16	T	1	T
colon	2000	–	–	T	28.263	+	T	0.188	32.57	T	4	**0.000**	0.00	T	3	1356.4
PCMAC	3289	–	–	T	96.930	+	T	0.094	19.11	T	5	**0.004**	0.76	T	1	T

For each dataset and solution method we report absolute (Gap$_{abs}$) and relative (Gap$_{rel}$) gaps, as well as time (Time, in seconds). We also report the number of iterations for the algorithms, denoted by #. For each dataset, among the exact solution methods, the best Time and the best Gap$_{abs}$ (if Time$\geq$ 3600 sec) are in **bold**, and p indicates the size of the full set of features

"–": No feasible solution is found within the time limit. "+": Gap is larger than 100. "T": Time limit (3600 sec.) is reached

Table 5 Classification performance of MILP$_1$, MILP$_2$ and MILP$_3$ when solving the mRMR feature selection problem (2); results for the full set of features are also reported

Dataset	Full set			MILP$_1$ (Chang 2001)			MILP$_2$ (Nguyen et al. 2009)			MILP$_3$		
	p	NB	RF	$\|S\|$	NB	RF	$\|S\|$	NB	RF	$\|S\|$	NB	RF
banknote_auth	4	0.84 ± 0.04	0.93 ± 0.03	4	**0.84** ± 0.04	**0.93** ± 0.03	4	**0.84** ± 0.04	**0.93** ± 0.03	4	**0.84** ± 0.04	**0.93** ± 0.03
Breast_cancer	9	0.68 ± 0.02	0.68 ± 0.03	7	**0.67** ± 0.03	**0.66** ± 0.02	7	**0.67** ± 0.03	**0.66** ± 0.02	7	**0.67** ± 0.03	**0.66** ± 0.02
Letter_Recog	16	0.35 ± 0.01	0.23 ± 0.00	14	**0.33** ± 0.01	**0.23** ± 0.00	14	**0.33** ± 0.01	**0.23** ± 0.00	14	**0.33** ± 0.01	**0.23** ± 0.00
Zoo	17	0.83 ± 0.09	0.85 ± 0.02	8	**0.77** ± 0.09	**0.83** ± 0.04	8	**0.77** ± 0.09	**0.83** ± 0.04	8	**0.77** ± 0.09	**0.83** ± 0.04
Breast_Cancer	31	0.49 ± 0.01	0.95 ± 0.02	22	**0.94** ± 0.02	**0.94** ± 0.02	22	**0.94** ± 0.02	**0.94** ± 0.02	22	**0.94** ± 0.02	**0.94** ± 0.02
SPECTF_Heart	44	0.72 ± 0.01	0.71 ± 0.01	5	**0.71** ± 0.01	**0.71** ± 0.01	5	**0.71** ± 0.01	**0.71** ± 0.01	5	**0.71** ± 0.01	**0.71** ± 0.01
Lung_Cancer	56	0.81 ± 0.05	0.69 ± 0.14	9	**0.60** ± 0.08	**0.85** ± 0.08	9	**0.60** ± 0.08	**0.85** ± 0.08	9	**0.60** ± 0.08	**0.85** ± 0.08
Sports_articles	59	0.82 ± 0.02	0.81 ± 0.02	1	**0.49** ± 0.00	**0.49** ± 0.00	2	**0.49** ± 0.00	**0.49** ± 0.00	1	**0.49** ± 0.00	**0.49** ± 0.00
Connectionist	60	0.66 ± 0.04	0.78 ± 0.06	39	**0.65** ± 0.05	0.73 ± 0.04	39	0.63 ± 0.04	**0.75** ± 0.06	44	0.63 ± 0.05	**0.75** ± 0.05
Optical_Recog	62	0.92 ± 0.01	0.82 ± 0.01	32	**0.91** ± 0.01	**0.79** ± 0.02	32	**0.91** ± 0.01	**0.79** ± 0.02	32	**0.91** ± 0.01	**0.79** ± 0.02
Hill-Valley	100	0.44 ± 0.02	0.51 ± 0.03	10	**0.44** ± 0.02	**0.50** ± 0.02	10	**0.44** ± 0.02	**0.50** ± 0.02	10	**0.44** ± 0.02	**0.50** ± 0.02
Urban_Land	147	0.75 ± 0.06	0.72 ± 0.07	90	0.79 ± 0.04	0.61 ± 0.08	42	0.81 ± 0.04	0.60 ± 0.07	29	**0.82** ± 0.04	**0.62** ± 0.05
Epileptic_Seiz	178	0.42 ± 0.01	0.28 ± 0.01	98	**0.42** ± 0.01	0.27 ± 0.01	91	**0.42** ± 0.01	0.27 ± 0.01	142	**0.42** ± 0.01	**0.28** ± 0.01
SCADI	205	0.78 ± 0.05	0.77 ± 0.05	18	**0.81** ± 0.07	**0.82** ± 0.04	15	0.79 ± 0.07	**0.82** ± 0.04	17	**0.81** ± 0.07	**0.82** ± 0.04
Semeion_Hand	256	0.85 ± 0.01	0.63 ± 0.03	25	0.63 ± 0.02	0.49 ± 0.03	26	0.62 ± 0.02	0.49 ± 0.03	42	**0.72** ± 0.01	**0.52** ± 0.03
USPS	256	0.78 ± 0.01	0.40 ± 0.01	30	0.46 ± 0.01	**0.40** ± 0.01	32	0.47 ± 0.01	0.38 ± 0.01	20	**0.48** ± 0.02	**0.40** ± 0.03
lung_discrete	325	0.77 ± 0.11	0.54 ± 0.13	17	0.46 ± 0.10	0.57 ± 0.09	33	0.76 ± 0.03	**0.66** ± 0.06	235	**0.84** ± 0.06	0.56 ± 0.07
Madelon	500	0.59 ± 0.01	0.61 ± 0.01	–	–	–	13	**0.59** ± 0.01	**0.61** ± 0.02	498	0.58 ± 0.01	**0.61** ± 0.02
ISOLET	617	0.84 ± 0.01	0.57 ± 0.01	1	0.04 ± 0.00	0.04 ± 0.00	44	0.56 ± 0.03	0.52 ± 0.02	617	**0.84** ± 0.01	**0.57** ± 0.01
Parkinson	754	0.73 ± 0.03	0.78 ± 0.03	–	–	–	487	0.72 ± 0.03	**0.78** ± 0.02	433	**0.73** ± 0.02	0.77 ± 0.02
CNAE-9	856	0.99 ± 0.00	1.00 ± 0.00	–	–	–	84	**1.00** ± 0.00	**1.00** ± 0.00	856	0.99 ± 0.00	**1.00** ± 0.00

Table 5 continued

Dataset	Full set			MILP$_1$ (Chang 2001)			MILP$_2$ (Nguyen et al. 2009)			MILP$_3$		
	p	NB	RF	$\|S\|$	NB	RF	$\|S\|$	NB	RF	$\|S\|$	NB	RF
Yale_32x32	1024	0.54 ± 0.08	0.43 ± 0.05	–	–	–	909	**0.55** ± 0.09	**0.51** ± 0.06	1024	0.54 ± 0.08	0.43 ± 0.05
ORL_32x32	1024	0.88 ± 0.02	0.48 ± 0.01	–	–	–	766	**0.89** ± 0.02	0.44 ± 0.04	748	**0.89** ± 0.03	**0.45** ± 0.05
colon	2000	0.67 ± 0.13	0.73 ± 0.06	–	–	–	150	**0.72** ± 0.11	**0.85** ± 0.10	2000	0.67 ± 0.13	0.73 ± 0.06
PCMAC	3289	0.92 ± 0.02	0.84 ± 0.03	–	–	–	3289	**0.92** ± 0.02	**0.84** ± 0.03	3289	**0.92** ± 0.02	**0.84** ± 0.03

For each dataset and solution method, we report the average ± standard deviation of F1 scores over the 5-fold cross validation results using Naive Bayes and Random Forest classifiers denoted by NB and RF, respectively. We also report the size of the full feature set (p) and the sizes of feature subsets ($|S|$) selected by the solution methods. For each dataset and classifier, among the solution methods, the best average F1 score is in **bold**

"–": No feasible solution is found within the time limit

Table 6 Classification performance of the best MILPs (MILP$_1$ and MILP$_3$), Algorithms 1 and 2, and the heuristic method (Heu.) when solving the mRMR feature selection problem (2); results for the full set of features are also reported

Dataset	Full set			MILP$_1$ (Chang 2001)			MILP$_3$			Algorithm 1		
	p	NB	RF	$\|S\|$	NB	RF	$\|S\|$	NB	RF	$\|S\|$	NB	RF
banknote_auth	4	0.84 ± 0.04	0.93 ± 0.03	4	**0.84** ± 0.04	**0.93** ± 0.03	4	**0.84** ± 0.04	**0.93** ± 0.03	4	**0.84** ± 0.04	**0.93** ± 0.03
Breast_cancer	9	0.68 ± 0.02	0.68 ± 0.03	7	**0.67** ± 0.03	0.66 ± 0.02	7	**0.67** ± 0.03	0.66 ± 0.02	7	**0.67** ± 0.03	0.66 ± 0.02
Letter_Recog	16	0.35 ± 0.01	0.23 ± 0.00	14	**0.33** ± 0.01	**0.23** ± 0.00	14	**0.33** ± 0.01	**0.23** ± 0.00	14	**0.33** ± 0.01	**0.23** ± 0.00
Zoo	17	0.83 ± 0.09	0.85 ± 0.02	8	0.77 ± 0.09	**0.83** ± 0.04	8	0.77 ± 0.09	**0.83** ± 0.04	8	0.77 ± 0.09	**0.83** ± 0.04
Breast_Cancer	31	0.49 ± 0.01	0.95 ± 0.02	22	**0.94** ± 0.02	0.94 ± 0.02	22	**0.94** ± 0.02	0.94 ± 0.02	25	0.93 ± 0.01	**0.95** ± 0.02
SPECTF_Heart	44	0.72 ± 0.01	0.71 ± 0.01	5	0.71 ± 0.01	**0.71** ± 0.01	5	0.71 ± 0.01	**0.71** ± 0.01	5	0.71 ± 0.01	**0.71** ± 0.01
Lung_Cancer	56	0.81 ± 0.05	0.69 ± 0.14	9	0.60 ± 0.08	**0.85** ± 0.08	9	0.60 ± 0.08	**0.85** ± 0.08	9	0.60 ± 0.08	**0.85** ± 0.08
Sports_articles	59	0.82 ± 0.02	0.81 ± 0.02	1	**0.49** ± 0.00	0.49 ± 0.00	1	**0.49** ± 0.00	0.49 ± 0.00	2	**0.49** ± 0.00	0.49 ± 0.00
Connectionist	60	0.66 ± 0.04	0.78 ± 0.06	39	**0.65** ± 0.05	0.73 ± 0.04	44	0.63 ± 0.05	0.75 ± 0.05	53	0.63 ± 0.04	**0.76** ± 0.07
Optical_Recog	62	0.92 ± 0.01	0.82 ± 0.01	32	0.91 ± 0.01	0.79 ± 0.02	32	0.91 ± 0.01	0.79 ± 0.02	32	0.91 ± 0.01	0.79 ± 0.02
Hill-Valley	100	0.44 ± 0.02	0.51 ± 0.03	10	**0.44** ± 0.02	0.50 ± 0.02	10	**0.44** ± 0.02	0.50 ± 0.02	10	**0.44** ± 0.02	0.50 ± 0.02
Urban_Land	147	0.75 ± 0.06	0.72 ± 0.07	90	0.79 ± 0.04	0.61 ± 0.08	29	**0.82** ± 0.04	0.62 ± 0.05	147	0.75 ± 0.06	**0.72** ± 0.07
Epileptic_Seiz	178	0.42 ± 0.01	0.28 ± 0.01	98	**0.42** ± 0.01	0.27 ± 0.01	142	**0.42** ± 0.01	**0.28** ± 0.01	80	**0.42** ± 0.01	0.27 ± 0.01
SCADI	205	0.78 ± 0.05	0.77 ± 0.05	18	**0.81** ± 0.07	**0.82** ± 0.04	17	**0.81** ± 0.07	**0.82** ± 0.04	14	0.79 ± 0.05	**0.82** ± 0.04
Semeion_Hand	256	0.85 ± 0.01	0.63 ± 0.03	25	0.63 ± 0.02	0.49 ± 0.03	42	0.72 ± 0.01	0.52 ± 0.03	25	0.61 ± 0.02	0.50 ± 0.03
USPS	256	0.78 ± 0.01	0.40 ± 0.01	30	0.46 ± 0.01	**0.40** ± 0.01	20	0.48 ± 0.02	**0.40** ± 0.03	40	0.45 ± 0.01	0.32 ± 0.01
lung_discrete	325	0.77 ± 0.11	0.54 ± 0.13	17	0.46 ± 0.10	0.57 ± 0.09	235	**0.84** ± 0.06	0.56 ± 0.07	33	0.78 ± 0.09	0.62 ± 0.03
Madelon	500	0.59 ± 0.01	0.61 ± 0.01	–	–	–	498	0.58 ± 0.01	**0.61** ± 0.02	500	**0.59** ± 0.01	**0.61** ± 0.01
ISOLET	617	0.84 ± 0.01	0.57 ± 0.01	1	0.04 ± 0.00	0.04 ± 0.00	617	**0.84** ± 0.01	**0.57** ± 0.01	617	**0.84** ± 0.01	**0.57** ± 0.01
Parkinson	754	0.73 ± 0.03	0.78 ± 0.03	–	–	–	433	**0.73** ± 0.02	0.77 ± 0.02	754	**0.73** ± 0.03	**0.78** ± 0.03
CNAE-9	856	0.99 ± 0.00	1.00 ± 0.00	–	–	–	856	0.99 ± 0.00	**1.00** ± 0.00	856	0.99 ± 0.00	**1.00** ± 0.00
Yale_32x32	1024	0.54 ± 0.08	0.43 ± 0.05	–	–	–	1024	**0.54** ± 0.08	**0.43** ± 0.05	1024	**0.54** ± 0.08	**0.43** ± 0.05

Table 6 continued

Dataset	Full set			MILP$_1$ (Chang 2001)			MILP$_3$			Algorithm 1		
	p	NB	RF	$\|S\|$	NB	RF	$\|S\|$	NB	RF	$\|S\|$	NB	RF
ORL_32x32	1024	0.88 ± 0.02	0.48 ± 0.01	–	–	–	748	**0.89** ± 0.03	0.45 ± 0.05	1024	0.88 ± 0.02	**0.48** ± 0.01
colon	2000	0.67 ± 0.13	0.73 ± 0.06	–	–	–	2000	0.67 ± 0.13	0.73 ± 0.06	2000	0.67 ± 0.13	0.73 ± 0.06
PCMAC	3289	0.92 ± 0.02	0.84 ± 0.03	–	–	–	3289	**0.92** ± 0.02	0.84 ± 0.03	3289	**0.92** ± 0.02	0.84 ± 0.03

Dataset	Algorithm 2			Heu. (Brown et al. 2012)		
	$\|S\|$	NB	RF	$\|S\|$	NB	RF
banknote_auth	4	**0.84** ± 0.04	**0.93** ± 0.03	2	**0.84** ± 0.03	0.85 ± 0.03
Breast_cancer	7	**0.67** ± 0.03	0.66 ± 0.02	2	0.65 ± 0.01	**0.68** ± 0.03
Letter_Recog	14	**0.33** ± 0.01	**0.23** ± 0.00	3	0.26 ± 0.00	**0.23** ± 0.00
Zoo	8	0.77 ± 0.09	**0.83** ± 0.04	18	**0.84** ± 0.08	**0.83** ± 0.03
Breast_Cancer	22	**0.94** ± 0.02	0.94 ± 0.02	2	0.49 ± 0.01	0.92 ± 0.02
SPECTF_Heart	5	0.71 ± 0.01	**0.71** ± 0.01	2	**0.72** ± 0.01	**0.71** ± 0.00
Lung_Cancer	9	0.60 ± 0.08	**0.85** ± 0.08	16	**0.90** ± 0.08	0.83 ± 0.11
Sports_articles	2	**0.49** ± 0.00	0.49 ± 0.00	4	**0.49** ± 0.00	**0.75** ± 0.01
Connectionist	39	0.64 ± 0.03	0.75 ± 0.05	2	0.52 ± 0.06	0.55 ± 0.05
Optical_Recog	32	0.91 ± 0.01	0.79 ± 0.02	63	**0.92** ± 0.01	**0.81** ± 0.01
Hill-Valley	10	**0.44** ± 0.02	0.50 ± 0.02	2	**0.44** ± 0.02	**0.52** ± 0.03
Urban_Land	84	0.80 ± 0.05	0.65 ± 0.08	2	0.64 ± 0.07	0.59 ± 0.08
Epileptic_Seiz	88	**0.42** ± 0.01	0.27 ± 0.01	2	0.34 ± 0.00	0.24 ± 0.01
SCADI	15	0.79 ± 0.07	0.80 ± 0.07	206	0.78 ± 0.05	0.72 ± 0.13
Semeion_Hand	26	0.61 ± 0.02	0.50 ± 0.03	257	**0.85** ± 0.01	**0.60** ± 0.02
USPS	71	**0.49** ± 0.03	0.32 ± 0.01	2	0.08 ± 0.01	0.17 ± 0.01
lung_discrete	32	0.78 ± 0.04	**0.66** ± 0.07	326	0.77 ± 0.11	0.55 ± 0.09

Table 6 continued

Dataset	Algorithm 2			Heu. (Brown et al. 2012)		
	$\lvert S\rvert$	NB	RF	$\lvert S\rvert$	NB	RF
Madelon	500	**0.59** ± 0.01	**0.61** ± 0.01	2	0.53 ± 0.03	0.49 ± 0.03
ISOLET	29	0.58 ± 0.01	0.37 ± 0.01	2	0.06 ± 0.01	0.07 ± 0.00
Parkinson	656	0.72 ± 0.02	**0.78** ± 0.03	2	0.64 ± 0.00	0.66 ± 0.02
CNAE-9	731	0.99 ± 0.00	**1.00** ± 0.00	20	**1.00** ± 0.00	**1.00** ± 0.00
Yale_32x32	1024	**0.54** ± 0.08	**0.43** ± 0.05	2	0.05 ± 0.03	0.11 ± 0.05
ORL_32x32	1024	0.88 ± 0.02	**0.48** ± 0.01	2	0.23 ± 0.04	0.14 ± 0.03
colon	2000	0.67 ± 0.13	0.73 ± 0.06	67	**0.78** ± 0.11	**0.86** ± 0.13
PCMAC	3289	**0.92** ± 0.02	0.84 ± 0.03	158	**0.92** ± 0.01	**0.86** ± 0.01

For each dataset and solution method, we report the average ± standard deviation of F1 scores over the 5-fold cross validation results using Naive Bayes and Random Forest classifiers denoted by NB and RF, respectively. We also report the size of the full feature set (p) and the sizes of feature subsets ($|S|$) selected by the solution methods. For each dataset and classifier, among the solution methods, the best average F1 score is in **bold**

"–": No feasible solution is found within the time limit

Table 7 Classification performance of MILP$_1$ and MILP$_2$, Algorithms 1 and 2, and the heuristic method (Heu.) when solving the CFS feature selection problem (4); results for the full feature set are also reported

Dataset	Full set			MILP$_1$ (Chang 2001)			MILP$_2$ (Nguyen et al. 2009)			Algorithm 1		
	p	NB	RF	$\|S\|$	NB	RF	$\|S\|$	NB	RF	$\|S\|$	NB	RF
banknote_auth	4	0.84 ± 0.04	0.93 ± 0.03	1	**0.85** ± 0.03	0.85 ± 0.04	1	**0.85** ± 0.03	0.85 ± 0.04	1	**0.85** ± 0.03	0.85 ± 0.04
Breast_cancer	9	0.68 ± 0.02	0.68 ± 0.03	2	0.65 ± 0.01	**0.73** ± 0.04	2	0.65 ± 0.01	**0.73** ± 0.04	2	0.65 ± 0.01	**0.73** ± 0.04
Letter_Recog	16	0.35 ± 0.01	0.23 ± 0.00	3	0.29 ± 0.00	**0.23** ± 0.00	3	0.29 ± 0.00	**0.23** ± 0.00	3	0.29 ± 0.00	**0.23** ± 0.00
Zoo	17	0.83 ± 0.09	0.85 ± 0.02	4	0.63 ± 0.03	0.77 ± 0.03	4	0.63 ± 0.03	0.77 ± 0.03	4	0.63 ± 0.03	0.77 ± 0.03
Breast_Cancer	31	0.49 ± 0.01	0.95 ± 0.02	1	**0.91** ± 0.02	0.91 ± 0.02	1	**0.91** ± 0.02	0.91 ± 0.02	1	**0.91** ± 0.02	0.91 ± 0.02
SPECTF_Heart	44	0.72 ± 0.01	0.71 ± 0.01	1	0.71 ± 0.01	**0.71** ± 0.00	1	0.71 ± 0.01	**0.71** ± 0.00	1	0.71 ± 0.01	**0.71** ± 0.00
Lung_Cancer	56	0.81 ± 0.05	0.69 ± 0.14	2	0.60 ± 0.08	**0.90** ± 0.08	2	0.60 ± 0.08	**0.90** ± 0.08	2	0.60 ± 0.08	**0.90** ± 0.08
Sports_articles	59	0.82 ± 0.02	0.81 ± 0.02	1	0.49 ± 0.00	0.75 ± 0.01	1	0.49 ± 0.00	0.75 ± 0.01	1	0.49 ± 0.00	0.75 ± 0.01
Connectionist	60	0.66 ± 0.04	0.78 ± 0.06	10	**0.69** ± 0.05	**0.68** ± 0.04	10	**0.69** ± 0.05	**0.68** ± 0.04	10	**0.69** ± 0.05	**0.68** ± 0.04
Optical_Recog	62	0.92 ± 0.01	0.82 ± 0.01	23	**0.90** ± 0.01	**0.80** ± 0.01	23	**0.90** ± 0.01	**0.80** ± 0.01	23	**0.90** ± 0.01	**0.80** ± 0.01
Hill-Valley	100	0.44 ± 0.02	0.51 ± 0.03	1	**0.45** ± 0.02	**0.50** ± 0.02	1	**0.45** ± 0.02	**0.50** ± 0.02	3	0.44 ± 0.02	0.49 ± 0.04
Urban_Land	147	0.75 ± 0.06	0.72 ± 0.07	7	**0.81** ± 0.05	**0.62** ± 0.06	7	**0.81** ± 0.05	**0.62** ± 0.06	9	0.77 ± 0.05	0.60 ± 0.05
Epileptic_Seiz	178	0.42 ± 0.01	0.28 ± 0.01	107	**0.42** ± 0.01	**0.27** ± 0.01	32	**0.42** ± 0.01	**0.27** ± 0.01	28	0.41 ± 0.01	**0.27** ± 0.01
SCADI	205	0.78 ± 0.05	0.77 ± 0.05	7	**0.80** ± 0.06	0.78 ± 0.02	7	0.74 ± 0.06	0.78 ± 0.08	7	0.74 ± 0.06	0.78 ± 0.08
Semeion_Hand	256	0.85 ± 0.01	0.63 ± 0.03	256	**0.85** ± 0.01	**0.63** ± 0.03	69	0.84 ± 0.01	**0.63** ± 0.02	256	**0.85** ± 0.01	**0.63** ± 0.03
USPS	256	0.78 ± 0.01	0.40 ± 0.01	256	**0.78** ± 0.01	0.40 ± 0.01	2	0.24 ± 0.01	0.34 ± 0.01	256	**0.78** ± 0.01	0.40 ± 0.01
lung_discrete	325	0.77 ± 0.11	0.54 ± 0.13	22	0.49 ± 0.02	0.52 ± 0.05	34	0.77 ± 0.09	0.55 ± 0.04	325	0.77 ± 0.11	0.54 ± 0.13
Madelon	500	0.59 ± 0.01	0.61 ± 0.01	1	**0.62** ± 0.02	0.62 ± 0.02	3	0.61 ± 0.01	**0.64** ± 0.02	3	0.61 ± 0.01	**0.64** ± 0.02

Table 7 continued

Dataset	Full set			MILP$_1$ (Chang 2001)			MILP$_2$ (Nguyen et al. 2009)			Algorithm 1		
	p	NB	RF	$\|S\|$	NB	RF	$\|S\|$	NB	RF	$\|S\|$	NB	RF
ISOLET	617	0.84 ± 0.01	0.57 ± 0.01	1	0.04 ± 0.00	0.03 ± 0.00	69	0.81 ± 0.02	0.55 ± 0.01	617	**0.84** ± 0.01	**0.57** ± 0.01
Parkinson	754	0.73 ± 0.03	0.78 ± 0.03	754	0.73 ± 0.03	0.78 ± 0.03	17	**0.82** ± 0.02	**0.79** ± 0.03	2	0.65 ± 0.02	0.64 ± 0.01
CNAE-9	856	0.99 ± 0.00	1.00 ± 0.00	–	–	–	2	**1.00** ± 0.00	**1.00** ± 0.00	2	**1.00** ± 0.00	**1.00** ± 0.00
Yale_32x32	1024	0.54 ± 0.08	0.43 ± 0.05	–	–	–	9	0.50 ± 0.15	0.40 ± 0.06	1024	0.54 ± 0.08	0.43 ± 0.05
ORL_32x32	1024	0.88 ± 0.02	0.48 ± 0.01	–	–	–	32	0.80 ± 0.03	**0.51** ± 0.05	1024	**0.88** ± 0.02	0.48 ± 0.01
colon	2000	0.67 ± 0.13	0.73 ± 0.06	–	–	–	2	0.53 ± 0.06	0.82 ± 0.11	2000	0.67 ± 0.13	0.73 ± 0.06
PCMAC	3289	0.92 ± 0.02	0.84 ± 0.03	–	–	–	22	0.88 ± 0.02	0.82 ± 0.03	3289	**0.92** ± 0.02	**0.84** ± 0.03

Dataset	Algorithm 2			Heu. (Zhao et al. 2010)		
	$\|S\|$	NB	RF	$\|S\|$	NB	RF
banknote_auth	1	**0.85** ± 0.03	0.85 ± 0.04	4	0.84 ± 0.04	**0.93** ± 0.03
Breast_cancer	2	0.65 ± 0.01	**0.73** ± 0.04	6	**0.68** ± 0.02	0.68 ± 0.02
Letter_Recog	3	0.29 ± 0.00	**0.23** ± 0.00	11	**0.34** ± 0.00	**0.23** ± 0.00
Zoo	4	0.63 ± 0.03	0.77 ± 0.03	14	**0.82** ± 0.10	**0.83** ± 0.03
Breast_Cancer	1	**0.91** ± 0.02	0.91 ± 0.02	8	0.49 ± 0.01	**0.94** ± 0.02
SPECTF_Heart	1	0.71 ± 0.01	**0.71** ± 0.00	25	**0.72** ± 0.01	**0.71** ± 0.00
Lung_Cancer	6	0.65 ± 0.12	**0.90** ± 0.08	12	**0.86** ± 0.13	0.83 ± 0.11
Sports_articles	1	0.49 ± 0.00	0.75 ± 0.01	16	**0.78** ± 0.03	**0.80** ± 0.01
Connectionist	10	**0.69** ± 0.05	**0.68** ± 0.04	6	0.60 ± 0.07	**0.68** ± 0.06
Optical_Recog	27	**0.90** ± 0.01	0.76 ± 0.01	37	**0.90** ± 0.01	0.74 ± 0.02
Hill-Valley	1	**0.45** ± 0.02	**0.50** ± 0.02	6	0.44 ± 0.02	**0.50** ± 0.03
Urban_Land	19	0.78 ± 0.04	**0.62** ± 0.04	7	0.65 ± 0.05	0.53 ± 0.04
Epileptic_Seiz	57	**0.42** ± 0.01	**0.27** ± 0.01	14	0.40 ± 0.00	**0.27** ± 0.01
SCADI	7	0.74 ± 0.06	0.78 ± 0.08	21	**0.80** ± 0.06	**0.81** ± 0.05

Table 7 continued

Dataset	Algorithm 2			Heu. (Zhao et al. 2010)		
	$\|S\|$	NB	RF	$\|S\|$	NB	RF
Semeion_Hand	81	0.84 ± 0.01	0.53 ± 0.02	26	0.74 ± 0.02	0.56 ± 0.04
USPS	38	0.55 ± 0.01	**0.41** ± 0.01	9	0.45 ± 0.02	0.39 ± 0.03
lung_discrete	40	0.70 ± 0.11	0.57 ± 0.07	65	**0.86** ± 0.03	**0.63** ± 0.09
Madelon	3	0.61 ± 0.01	**0.64** ± 0.02	8	0.61 ± 0.01	0.63 ± 0.01
ISOLET	617	**0.84** ± 0.01	**0.57** ± 0.01	9	0.31 ± 0.02	0.30 ± 0.01
Parkinson	425	0.73 ± 0.03	0.75 ± 0.02	6	0.66 ± 0.01	0.64 ± 0.00
CNAE-9	2	**1.00** ± 0.00	**1.00** ± 0.00	11	**1.00** ± 0.00	**1.00** ± 0.00
Yale_32x32	697	**0.55** ± 0.08	**0.50** ± 0.08	21	0.50 ± 0.05	0.44 ± 0.09
ORL_32x32	1024	**0.88** ± 0.02	0.48 ± 0.01	21	0.63 ± 0.05	0.35 ± 0.02
colon	568	0.73 ± 0.08	0.81 ± 0.12	25	**0.82** ± 0.08	**0.88** ± 0.09
PCMAC	3289	**0.92** ± 0.02	**0.84** ± 0.03	10	0.85 ± 0.02	0.81 ± 0.02

For each dataset and solution method, we report the average ± standard deviation of F1 scores over the 5-fold cross validation results using Naive Bayes and Random Forest classifiers denoted by NB and RF, respectively. We also report the size of the full feature set (p) and the sizes of feature subsets ($|S|$) selected by the solution methods. For each dataset and classifier, among the solution methods, the best average F1 score is in **bold**

"–": No feasible solution is found within the time limit

Next, we compare the performance of the best two MILPs (i.e., MILP$_1$ and MILP$_3$ based on the above discussion) against Algorithms 1 and 2 as well as the heuristic method (Brown et al. 2012) when solving the mRMR problem; see Table 3. Parametric algorithms are competitive, if not better than pure MILP methods with $p \leq 60$. More importantly, they are able to solve problems with larger datasets to optimality, e.g., "Semeion_Handwritten_Digit" ($p = 256$) or "Madelon" ($p = 500$), and consistently report solutions with much smaller gaps than MILPs. In some cases, e.g., "Optical_Recognition" dataset ($p = 62$), the parametric algorithms prove optimality and are faster than the heuristic.

In Table 4, we report the computational results for the CFS feature selection problem. Similar to the aforementioned results for mRMR, note that for CFS the parametric algorithms outperform both MILP$_1$ and MILP$_2$ and scale better for larger instances. Moreover, for the CFS problem, the heuristic is, in fact, quite expensive: in all cases where either Algorithm 1 or 2 find optimal solutions, their runtime is a fraction of the heuristic runtime.

By comparing the computational performances of the parametric algorithms (Tables 3 and 4), we note that Algorithms 1 and 2 have similar running times for the datasets that they solve to optimality. For the datasets where an optimal solution is not found within the time limit, Algorithm 1 can be a better choice as for these datasets $\texttt{Gap}_{rel}$ and $\texttt{Gap}_{abs}$ reported by Algorithm 2 are approximations of the relative and absolute gaps, respectively.

In Table 5, we compare F1 classification score of the feature subsets obtained by MILPs when solving the mRMR problem (2). In most of the datasets, MILP$_3$ has the best score for the both Naive Bayes and Random Forest classifiers and for a few other datasets it has a competitive performance with the best reported scores. The mRMR heuristic, in general, has a worse classification performance than the exact methods for the both classifiers, see, e.g., datasets "Connectionist", "ISOLET" and "USPS" in Table 6. The exact methods also generally outperform the CFS heuristic method, in particular, for larger datasets ($p > 60$); see the results in Table 7.

On average, the parametric algorithms have similar classification performances. However, when solving the mRMR problem Algorithm 1 provides better results than Algorithm 2 for more datasets; see Table 6. Table 7 shows the opposite observation for the CFS problem. In Tables 6 and 7, there are some cases, where the MILPs report slightly better classification performances than the parametric methods; however, considering the fact that the parametric algorithms are faster, they can be used as the recommended solution methods for solving both the mRMR and CFS feature selection problems.

Finally, it is worth mentioning that the choice of an appropriate feature selection measure may depend on the dataset and its application setting [see, e.g., Chandrashekar and Sahin (2014) and Jović et al. (2015) for a comprehensive discussions]. In particular, due to the different structures and also coefficients values of the problems, the sizes of the selected subsets of features by CFS are typically smaller than those selected by mRMR.

6 Concluding remarks

Feature selection is an essential preprocessing step in many data mining and machine learning tasks and involves finding a small subset of the most relevant features from the dataset. In this paper, we focus on feature selection problems based on mRMR and CFS measures that are typically tackled either by heuristic methods or their reformulations as MILPs. However, heuristics do not guarantee the optimality of the output feature subset and MILPs given in the literature have rather poor performances even for small- and medium-sized datasets.

To address the aforementioned shortcomings, we consider approaches that ensure globally optimal solutions. To this end, we propose another MILP reformulation for the mRMR feature selection problem which outperforms existing MILPs in the literature. Additionally, we apply parametric approaches from fractional optimization to solve both the mRMR and CFS feature selection problems. Our computational experiments with real-world datasets show that the proposed approaches lead to encouraging improvements.

Acknowledgements The authors thank the Review Team for their helpful and constructive comments. This paper is based upon work supported by the National Science Foundation under Grant No. 1818700.

References

Ahuja, R. K., Magnanti, T. L., & Orlin, J. B. (1993). *Network flows: Theory, algorithms, and applications*. London: Pearson Education.

Asuncion, A. & Newman, D. (2007). UCI machine learning repository. https://archive.ics.uci.edu. Accessed August 2020.

Atamtürk, A. & Gómez, A. (2020). Safe screening rules for l0-regression from perspective relaxations. In *International conference on machine learning* (pp. 421–430). PMLR.

Borrero, J. S., Gillen, C., & Prokopyev, O. A. (2017). Fractional 0–1 programming: Applications and algorithms. *Journal of Global Optimization*, *69*(1), 255–282.

Brown, G., Pocock, A., Zhao, M.-J., & Luján, M. (2012). Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *The Journal of Machine Learning Research*, *13*(1), 27–66.

Busygin, S., Prokopyev, O. A., & Pardalos, P. M. (2005). Feature selection for consistent biclustering via fractional 0–1 programming. *Journal of Combinatorial Optimization*, *10*(1), 7–21.

Busygin, S., Prokopyev, O., & Pardalos, P. M. (2008). Biclustering in data mining. *Computers & Operations Research*, *35*(9), 2964–2987.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16–28.

Chang, C.-T. (2001). On the polynomial mixed 0–1 fractional programming problems. *European Journal of Operational Research*, *131*(1), 224–227.

Cilia, N. D., De Stefano, C., Fontanella, F., & di Freca, A. S. (2019). A ranking-based feature selection approach for handwritten character recognition. *Pattern Recognition Letters*, *121*, 77–86.

Ding, C., & Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology*, *3*(02), 185–205.

Dinkelbach, W. (1967). On nonlinear fractional programming. *Management Science*, *13*(7), 492–498.

El Ghaoui, L., Viallon, V., & Rabbani, T. (2010). Safe feature elimination for the lasso and sparse supervised learning problems. arXiv preprint arXiv:1009.4219.

Fan, Y.-J., & Chaovalitwongse, W. A. (2010). Optimizing feature selection to improve medical diagnosis. *Annals of Operations Research*, *174*(1), 169–183.

Glover, F., & Woolsey, E. (1974). Converting the 0–1 polynomial programming problem to a 0–1 linear program. *Operations Research*, *22*(1), 180–182.

Gómez, A., & Prokopyev, O. A. (2020). A mixed-integer fractional optimization approach to best subset selection. INFORMS Journal on Computing. Accepted for publication.

Gurobi (2018). Gurobi optimizer reference manual v. 8. http://www.gurobi.com. Accessed August 2020.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*(Mar), 1157–1182.

Hall, M. A. (1999). *Correlation-based feature selection for machine learning*. PhD thesis, University of Waikato Hamilton.

Huang, H., Xie, H.-B., Guo, J.-Y., & Chen, H.-J. (2012). Ant colony optimization-based feature selection method for surface electromyography signals classification. *Computers in Biology and Medicine*, *42*(1), 30–38.

Ibaraki, T. (1983). Parametric approaches to fractional programs. *Mathematical Programming*, *26*(3), 345–362.

IBM (2019). ILOG CPLEX Optimizer v. 12.9.0. http://www-01.ibm.com. Accessed August 2020.

Jović, A., Brkić, K., & Bogunović, N. (2015). A review of feature selection methods with applications. In *2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)*,pp. 1200–1205. IEEE.

Kocheturov, A., Pardalos, P. M., & Karakitsiou, A. (2019). Massive datasets and machine learning for computational biomedicine: Trends and challenges. *Annals of Operations Research*, *276*(1–2), 5–34.

Lawler, E. L. (2001). *Combinatorial optimization: Networks and matroids*. North Chelmsford: Courier Corporation.

Li, J., Cheng, K., Wang, S., Morstatter, F., Robert, T., Tang, J., & Liu, H. (2016). Feature selection: A data perspective. arXiv:1601.07996.

Liu, H., & Motoda, H. (2012). *Feature selection for knowledge discovery and data mining* (Vol. 454). Berlin: Springer.

MacKay, D. J. C. (2003). *Information theory, inference and learning algorithms*. Cambridge: Cambridge University Press.

Megiddo, N. (1979). Combinatorial optimization with rational objective functions. *Mathematics of Operations Research*, *4*(4), 414–424.

Mehmanchi, E. (2020). *Reformulation Techniques and Solution Approaches for Fractional 0-1 Programs and Applications*. PhD thesis, University of Pittsburgh.

Mehmanchi, E., Gómez, A., & Prokopyev, O. A. (2019). Fractional 0–1 programs: Links between mixed-integer linear and conic quadratic formulations. *Journal of Global Optimization*, *75*(2), 273–339.

Nguyen, H. T., Franke, K. & Petrović, S. (2011). A new ensemble-feature-selection framework for intrusion detection. In *2011 11th international conference on intelligent systems design and applications*, pages 213–218. IEEE.

Nguyen, H., Franke, K., & Petrovic, S. (2010a). Improving effectiveness of intrusion detection by correlation feature selection. In *2010 International conference on availability, reliability and security* (pp. 17–24). IEEE.

Nguyen, H. T., Franke, K., & Petrovic, S. (2010b). Towards a generic feature-selection measure for intrusion detection. In *2010 20th international conference on pattern recognition* (pp. 1529–1532). IEEE.

Nguyen, H., Franke, K.,& Petrovic, S. (2009). Optimizing a class of feature selection measures. In NIPS 2009 *Workshop on discrete optimization in machine learning: Submodularity, Sparsity & Polyhedra (DISCML)*. Canada: Vancouver.

Palubeckis, G. (2004). Multistart tabu search strategies for the unconstrained binary quadratic optimization problem. *Annals of Operations Research*, *131*(1–4), 259–282.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

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.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). Numerical recipes in C++. *The Art of Scientific Computing*, *2*, 1002.

Python Software Foundation. (2020). Python language reference v. 3.7.7. https://www.python.org/. Last accessed August (2020).

Radzik, T. (2013). Fractional combinatorial optimization. In *Handbook of combinatorial optimization* (pp. 1311–1355). Springer.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507–2517.

Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review. *Data Classification: Algorithms and Applications*, p. 37.

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., et al. (2012). Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *74*(2), 245–266.

Viola, M., Sangiovanni, M., Toraldo, G., & Guarracino, M. R. (2017). A generalized eigenvalues classifier with embedded feature selection. *Optimization Letters*, *11*(2), 299–311.

Yu, L. & Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th international conference on machine learning (ICML-03)* (pp. 856–863).

Yuan, H., Xu, W., Li, Q., & Lau, R. (2018). Topic sentiment mining for sales performance prediction in e-commerce. *Annals of Operations Research*, *270*(1–2), 553–576.

Zhao, Z., Morstatter, F., Sharma, S., Alelyani, S., Anand, A., & Liu, H. (2010). Advancing feature selection research. ASU feature selection repository (pp. 1–28).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.