Machine Learning
https://doi.org/10.1007/s10994-018-5728-y

An adaptive heuristic for feature selection based on complementarity

Sumanta Singha[1] · **Prakash P. Shenoy**[2]

Received: 17 July 2017 / Accepted: 12 June 2018

Abstract
Feature selection is a dimensionality reduction technique that helps to improve data visualization, simplify learning, and enhance the efficiency of learning algorithms. The existing redundancy-based approach, which relies on relevance and redundancy criteria, does not account for feature complementarity. Complementarity implies information synergy, in which additional class information becomes available due to feature interaction. We propose a novel filter-based approach to feature selection that explicitly characterizes and uses feature complementarity in the search process. Using theories from multi-objective optimization, the proposed heuristic penalizes redundancy and rewards complementarity, thus improving over the redundancy-based approach that penalizes all feature dependencies. Our proposed heuristic uses an adaptive cost function that uses redundancy–complementarity ratio to automatically update the trade-off rule between relevance, redundancy, and complementarity. We show that this adaptive approach outperforms many existing feature selection methods using benchmark datasets.

Keywords Dimensionality reduction · Feature selection · Classification · Feature complementarity · Adaptive heuristic

1 Introduction

Learning from data is one of the central goals of machine learning research. Statistical and data-mining communities have long focused on building simpler and more interpretable models for prediction and understanding of data. However, high dimensional data present unique computational challenges such as model over-fitting, computational intractability,

Editor: Gavin Brown.

✉ Sumanta Singha
sumanta_singha@isb.edu

Prakash P. Shenoy
pshenoy@ku.edu

[1] Indian School of Business (ISB), Hyderabad, India

[2] School of Business, University of Kansas, Kansas, USA

Published online: 19 June 2018

and poor prediction. Feature selection is a dimensionality reduction technique that helps to simplify learning, reduce cost, and improve data interpretability.

Existing approaches Over the years, feature selection methods have evolved from simplest univariate ranking algorithms to more sophisticated relevance-redundancy trade-off to interaction-based approach in recent times. Univariate feature ranking (Lewis 1992) is a feature selection approach that ranks features based on relevance and ignores redundancy. As a result, when features are interdependent, the ranking approach leads to sub-optimal results (Brown et al. 2012). Redundancy-based approach improves over the ranking approach by considering both relevance and redundancy in the feature selection process. A wide variety of feature selection methods are based on relevance-redundancy trade-off (Battiti 1994; Hall 2000; Yu and Liu 2004; Peng et al. 2005; Ding and Peng 2005; Senawi et al. 2017). Their goal is to find an optimal subset of features that produces maximum relevance and minimum redundancy.

Complementarity-based feature selection methods emerged as an alternative approach to account for feature complementarity in the selection process. Complementarity can be described as a phenomenon in which two features together provide more information about the target variable than the sum of their individual information (information synergy). Several complementarity-based methods are proposed in the literature (Yang and Moody 1999, 2000; Zhao and Liu 2007; Meyer et al. 2008; Bontempi and Meyer 2010; Zeng et al. 2015; Chernbumroong et al. 2015). Yang and Moody (1999) and later Meyer et al. (2008) propose an interactive sequential feature selection method, known as joint mutual information (JMI), which selects a candidate feature that maximizes relevance and complementarity simultaneously. They conclude that JMI approach provides the best trade-off in terms of accuracy, stability, and flexibility with small data samples.

Limitations of the existing methods Clearly, redundancy-based approach is less efficient than complementarity-based approach as it does not account for feature complementarity. However, the main criticism of redundancy-based approach is with regard to how redundancy is formalized and measured. Correlation is the most common way to measure redundancy between features, which implicitly assumes that all correlated features are redundant. However, Guyon and Elisseeff (2003), Gheyas and Smith (2010) and Brown et al. (2012) show that this is an incorrect assumption; correlation does not imply redundancy, nor absence of complementarity. This is evident in Figs. 1 and 2, which present a 2-class classification problem (denoted by star and circle) with two continuous features X_1 and X_2. The projections on the axis denote the relevance of each respective feature. In Fig. 1, X_1 and X_2 are perfectly correlated, and indeed redundant. Having both X_1 and X_2 leads to no significant improvement in class separation compared to having either X_1 or X_2. However, in Fig. 2, a perfect separation is achieved by X_1 and X_2 together, although they are (negatively) correlated (within each class), and have identical relevance as in Fig. 1. This shows that while generic dependency is undesirable, dependency that conveys class information is useful. Whether two interacting features are redundant or complimentary depends on the relative magnitude of their class conditional dependency in comparison to their unconditional dependency. However, in the redundancy-based approach, it only focuses on unconditional dependency, and try to minimize it without exploring whether such dependencies lead to information gain or loss.

Although complementarity-based approach overcomes some of the drawbacks of redundancy based approach, it has limitations, too. Most of the existing complementarity-based feature selection methods either adopt a sequential feature selection approach or evaluate subsets of a given size. Sequential feature selection method raises a few important issues. First, Zhang and Zhang (2012) show that sequential selection methods suffer from initial

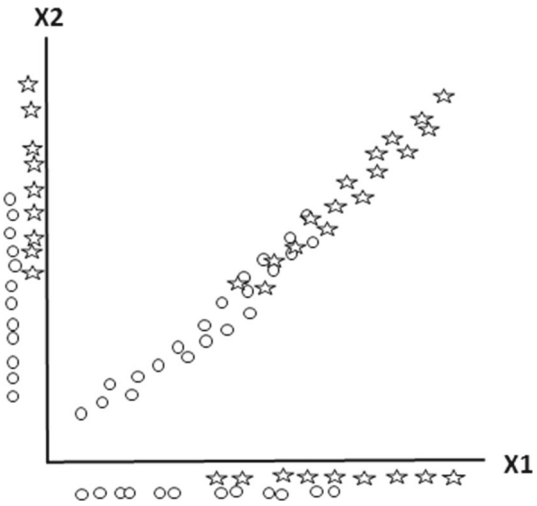

Fig. 1 Perfectly correlated and redundant

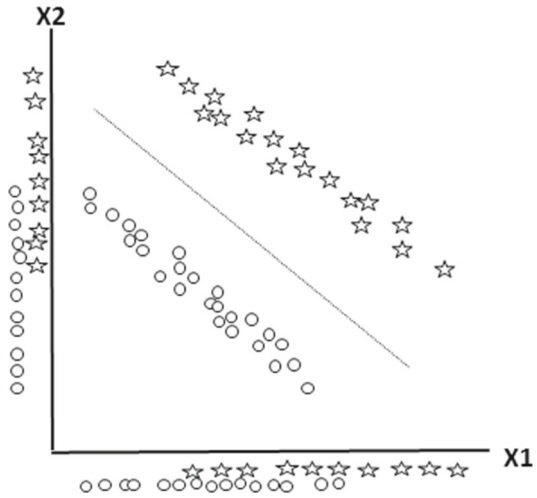

Fig. 2 Features are negatively-correlated within class, yet provide perfect separation

selection bias, i.e., the feature selected in the earlier steps govern acceptance or rejection of the subsequent features at each iteration and not vice versa. Second, it requires a priori knowledge of the desired number of features to be selected or some stopping criterion, which is usually determined by expert information or some technical considerations such as scalability or computation time. In many practical situations, it is difficult to get such prior knowledge. Moreover, in many cases, we are interested in finding an optimal feature subset that gives maximum predictive accuracy for a given task, and we are not really concerned about their ranking or the size of the optimal subset. Third, most of the existing methods combine the redundancy and complementarity and consider the net or aggregate effect in the search process (note that complementarity has the opposite sign of redundancy). Bontempi and Meyer (2010) refer to this aggregate approach as 'implicit' consideration of complementarity.

Our approach In this paper, we propose a filter-based feature subset selection method based on relevance, redundancy, and complementarity. Unlike most of the existing methods, which

focus on feature ranking or compare subsets of a given size, our goal is to select an optimal subset of features that predicts well. This is useful in many situations, where no prior expert knowledge is available regarding size of an optimal subset or the goal is simply to find an optimal subset. Using a multi-objective optimization (MOO) technique and an adaptive cost function, the proposed method aims to (1) maximize relevance, (2) minimize redundancy, and (3) maximize complementarity, while keeping the subset size as small as possible.

The term 'adaptive' implies that our proposed method adaptively determines the trade-off between relevance, redundancy, and complementarity based on subset properties. An adaptive approach helps to overcome the limitations of a fixed policy that fails to model the trade-off between competing objectives appropriately in a MOO problem. Such an adaptive approach is new to feature selection and essentially mimics a feature feedback mechanism in which the trade-off rule is a function of the objective values. The proposed cost function is also flexible in that it does not assume any particular functional form or rely on concavity assumption, and uses implicit utility maximization principles (Roy 1971; Rosenthal 1985).

Unlike some of the complementarity-based methods, which consider the net (aggregate) effect of redundancy and complementarity, we consider 'redundancy minimization' and 'complementarity maximization' as two separate objectives in the optimization process. This allows us the flexibility to apply different weights (preference) to redundancy and complementarity and control their relative importance adaptively during the search process. Using best-first as search strategy, the proposed heuristic offers a "best compromise" solution (more likely to avoid local optimum due to interactively determined gradient), if not the "best solution (in the sense of optimum)" (Saska 1968), which is sufficiently good in most practical scenarios. Using benchmark datasets, we show empirically that our adaptive heuristic not only outperforms many redundancy-based methods, but is also competitive amongst the existing complementarity-based methods.

Structure of the paper The rest of the paper is organized as follows. Section 2 presents the information-theoretic definitions and the concepts of relevance, redundancy, and complementarity. In Sect. 3, we present the existing feature selection methods, and discuss their strengths and limitations. In Sect. 4, we describe the proposed heuristic, and its theoretical motivation. In this section, we also discuss the limitations of our heuristic, and carry out sensitivity analysis. Section 5 presents the algorithm for our proposed heuristic, and evaluates its time complexity. In Sect. 6, we assess the performance of the heuristic on two synthetic datasets. In Sect. 7, we validate our heuristic using real data sets, and present the experimental results. In Sect. 8, we summarize and conclude.

2 Information theory: definitions and concepts

First, we provide the necessary definitions in information theory (Cover and Thomas 2006) and then discuss the existing notions of relevance, redundancy, and complementarity.

2.1 Definitions

Suppose, X and Y are discrete random variables with finite state spaces $\mathscr{X}$ and $\mathscr{Y}$, respectively. Let $p_{X,Y}$ denote the joint probability mass function (PMF) of X and Y, with marginal PMFs p_X and p_Y.

Definition 1 (*Entropy*) Entropy of X, denoted by $H(X)$, is defined as follows: $H(X) = -\sum_{x\in\mathscr{X}} p_X(x)\log(p_X(x))$. Entropy is a measure of uncertainty in PMF p_X of X.

Definition 2 (*Joint entropy*) Joint entropy of X and Y, denoted by $H(X, Y)$, is defined as follows: $H(X, Y) = -\sum_{x\in\mathscr{X}}\sum_{y\in\mathscr{Y}} p_{X,Y}(x, y)\log(p_{X,Y}(x, y))$. Joint entropy is a measure of uncertainty in the joint PMF $p_{X,Y}$ of X and Y.

Definition 3 (*Conditional entropy*) Conditional entropy of X given Y, denoted by $H(X|Y)$, is defined as follows: $H(X|Y) = -\sum_{x\in\mathscr{X}}\sum_{y\in\mathscr{Y}} p_{X,Y}(x, y)\log(p_{X|y}(x))$, where $p_{X|y}(x)$ is the conditional PMF of X given $Y = y$. Conditional entropy $H(X|Y)$ measures the remaining uncertainty in X given the knowledge of Y.

Definition 4 (*Mutual information (MI)*) Mutual information between X and Y, denoted by $I(X; Y)$, is defined as follows: $I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$. Mutual information measures the amount of dependence between X and Y. It is non-negative, symmetric, and is equal to zero *iff* X and Y are independent.

Definition 5 (*Conditional mutual information*) Conditional mutual information between X and Y given another discrete random variable Z, denoted by $I(X; Y|Z)$, is defined as follows: $I(X; Y|Z) = H(X|Z) - H(X|Y, Z) = H(Y|Z) - H(Y|X, Z)$. It measures the conditional dependence between X and Y given Z.

Definition 6 (*Interaction information*) Interaction information (McGill 1954; Matsuda 2000; Yeung 1991) among X, Y, and Z, denoted by $I(X; Y; Z)$, is defined as follows: $I(X; Y; Z) = I(X; Y) - I(X; Y|Z)$.[1] Interaction information measures the change in the degree of association between two random variables by holding one of the interacting variables constant. It can be positive, negative, or zero depending on the relative order of magnitude of $I(X; Y)$ and $I(X; Y|Z)$. Interaction information is symmetric (order independent). More generally, the interaction information among a set of n random variables $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$, is given by $I(X_1; X_2; \ldots; X_n) = -\sum_{\mathbf{S}\in\mathbf{X}'} (-1)^{|\mathbf{S}|} H(\mathbf{S})$ where, $\mathbf{X}'$ is the superset of $\mathbf{X}$ and $\sum$ denotes the sum over all subsets $\mathbf{S}$ of the superset $\mathbf{X}'$ (Abramson 1963). Should it be zero, we say that features do not interact 'altogether' (Kojadinovic 2005).

Definition 7 (*Multivariate mutual information*) Multivariate mutual information (Kojadinovic 2005; Matsuda 2000) between a set of n features $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ and Y, denoted as follows: $I(\mathbf{X}; Y)$, is defined by $I(\mathbf{X}; Y) = \sum_i I(X_i; Y) - \sum_{i<j} I(X_i; X_j; Y) + \cdots + (-1)^{n+1} I(X_1; \ldots; X_n; Y)$. This is the möbius representation of multivariate mutual information based on set theory. Multivariate mutual information measures the information that $\mathbf{X}$ contains about Y and can be seen as a series of alternative inclusion and exclusion of higher-order terms that represent the simultaneous interaction of several variables.

[1] In our paper, we use $I(X; Y; Z) = I(X; Y) - I(X; Y|Z)$ as interaction information that uses the sign convention consistent with the measure theory and is used by several authors Meyer and Bontempi (2006), Meyer et al. (2008) and Bontempi and Meyer (2010). Jakulin and Bratko (2004) defines $I(X; Y; Z) = I(X; Y|Z) - I(X; Y)$ as interaction information, which has opposite signs for odd number of random variables. Either formulation measures the same aspect of feature interaction (Krippendorff 2009). The sign convention used in this paper corresponds to the common area of overlap in the information diagram and does not impact the heuristic as we deal with absolute value of the interaction information.

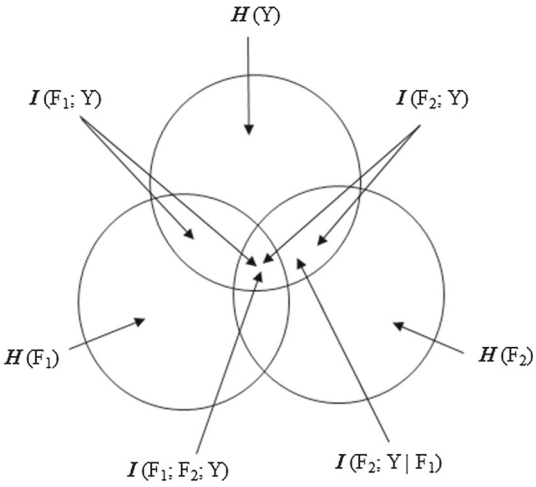

Fig. 3 Venn diagram showing the interaction between features F_1, F_2, and class Y

Definition 8 (*Symmetric uncertainty*) Symmetric uncertainty (Witten et al. 2016) between X and Y, denoted by $SU(X, Y)$, is defined as follows: $SU(X, Y) = \frac{2\,I(X;Y)}{H(X)+H(Y)}$. Symmetric uncertainty is a normalized version of MI in the range [0, 1]. Symmetric uncertainty can compensate for MI's bias towards features with more values.

Definition 9 (*Conditional symmetric uncertainty*) Conditional symmetric uncertainty between X and Y given Z, denoted by $SU(X, Y|Z)$, is defined as follows: $SU(X, Y|Z) = \frac{2\,I(X;Y|Z)}{H(X|Z)+H(Y|Z)}$. $SU(X, Y|Z)$ is a normalized version of conditional mutual information $I(X; Y|Z)$.

From Definitions 8 and 9, the symmetric uncertainty equivalent of interaction information can be expressed as follows: $SU(X, Y, Z) = SU(X, Y) - SU(X, Y|Z)$. Using the above notations, we can formulate the feature selection problem as follows: Given a set of n features $\mathbf{F} = \{F_i\}_{i\in\{1,\ldots,n\}}$, the goal of feature selection to select a subset of features $\mathbf{F}_S = \{F_i : i \in S\}, S \subseteq \{1, 2, \ldots, n\}$ such that $\mathbf{F}_S = \arg\max_S I(\mathbf{F}_S; Y)$, where $I(\mathbf{F}_S; Y)$ denotes mutual information between F_S and the class variable Y. For tractability reasons and unless there is strong evidence for the existence of higher-order interaction, the correction terms beyond 3-way interaction are generally ignored in the estimation of multivariate mutual information. In this paper, we will use

$$I(\mathbf{F}_S; Y) \approx \sum_{i\in S} I(F_i; Y) - \sum_{i,j\in S, i<j} I(F_i; F_j; Y) \tag{1}$$

where $I(F_i; F_j; Y)$ is the 3-way interaction term between F_i, F_j, and Y. The proof of Eq. 1 for a 3-variable case can be shown easily using Venn diagram shown in Fig. 3. The n variable case can be shown by recursive computation of the 3-variable case.

$$\begin{aligned} I(F_1, F_2; Y) &= I(F_1; Y) + I(F_2; Y|F_1) \\ &= I(F_1; Y) + I(F_2; Y) - I(F_1; F_2; Y) \end{aligned} \tag{2}$$

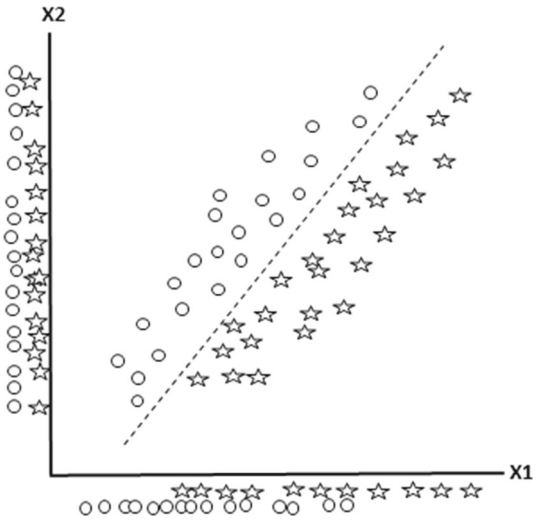

Fig. 4 X_2 is individually irrelevant but improves the separability of X_1

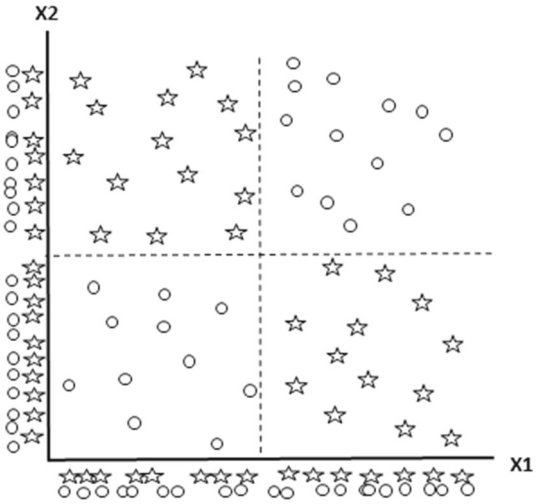

Fig. 5 Both individually irrelevant features become relevant together

2.2 Relevance

Relevance of a feature signifies its explanatory power, and is a measure of feature worthiness. A feature can be relevant individually or together with other variables if it carries information about the class Y. It is also possible that an individually relevant feature becomes relevant or a relevant feature becomes irrelevant when other features are present. This can be shown using Figs. 4 and 5, which present a 2-class classification problem (denoted by star and circle) with two continuous features X_1 and X_2. The projections of the class on each axis denotes each feature's individual relevance. In Fig. 4, X_2, which is individually irrelevant (uninformative), becomes relevant in the presence of X_1 and together improve the class separation that is otherwise achievable by X_1 alone. In Fig. 5, both X_1 and X_2 are individually irrelevant, however provide a perfect separation when present together ("chessboard problem," analogous to XOR problem). Thus relevance of a feature is context dependent (Guyon and Elisseeff 2003).

Using information-theoretic framework, a feature F_i is said to be unconditionally relevant to the class variable Y if $I(F_i; Y) > 0$, and irrelevant if $I(F_i; Y) = 0$. When evaluated in the context of other features, we call F_i to be conditionally relevant if $I(F_i; Y|F_{S-i}) > 0$, where $F_{S-i} = F_S \setminus F_i$. There are several other probabilistic definitions of relevance available in the literature. Most notably, Kohavi and John (1997) formalize relevance in terms of an optimal Bayes classifier and propose 2° of relevance—strong and weak. Strongly relevant features are those that bring unique information about the class variable and can not be replaced by other features. Weakly relevant features are relevant but not unique in the sense that they can be replaced by other features. An irrelevant feature is one that is neither strong nor weakly relevant.

2.3 Redundancy

The concept of redundancy is associated with the degree of dependency between two or more features. Two variables are said to be redundant, if they share common information about each other. This is the general dependency measured by $I(F_i; F_j)$. McGill (1954) and Jakulin and Bratko (2003) formalize this notion of redundancy as multi-information or total correlation. The multi-information between a set of n features $\{F_1, \ldots, F_n\}$ is given by $R(F_1, \ldots, F_n) = \sum_{i=1}^{n} H(F_i) - H(F_1, \ldots, F_n)$. For $n = 2$, $R(F_1, F_2) = H(F_1) + H(F_2) - H(F_1, F_2) = I(F_1; F_2)$. This measure of redundancy is non-linear, non-negative and non-decreasing with the number of features. In the context of feature election, it is often of interest to know whether two features are redundant with respect to the class variable, more than whether they are mutually redundant. Two features F_i and F_j are said to be redundant with respect to the class variable Y, if $I(F_i, F_j; Y) < I(F_i; Y) + I(F_j; Y)$. From Eq. 2, it follows $I(F_i; F_j; Y) > 0$ or $I(F_i; F_j) > I(F_i; F_j|Y)$. Thus two features are redundant with respect to the class variable if their unconditional dependency exceeds their class-conditional dependency.

2.4 Complementarity

Complementarity, known as information synergy, is the beneficial effect of feature interaction where two features together provide more information than the sum of their individual information. Two features F_i and F_j are said to be complementary with respect to the class variable Y if $I(F_i, F_j; Y) > I(F_i; Y) + I(F_j; Y)$, or equivalently, $I(F_i; F_j) < I(F_i; F_j|Y)$. Complementarity is negative interaction information. While generic dependency is undesirable, the dependency that conveys class information is good. Different researchers have explained complementarity from different perspectives. Vergara and Estévez (2014) define complementarity in terms of the degree of interaction between an individual feature F_i and the selected feature subset F_S given the class Y, i.e., $I(F_i, F_S|Y)$. Brown et al. (2012) provide similar definition to complementarity but call it conditional redundancy. They come to the similar conclusion as (Guyon and Elisseeff 2003): 'the inclusion of the correlated features can be useful, provided the correlation within the class is stronger than the overall correlation.'

3 Related literature

In this section, we review filter-based feature selection methods, which use information gain as a measure of dependence. In terms of evaluation strategy, filter-based methods can be broadly

classified into (1) redundancy-based approach, and (2) complementarity-based approach depending on whether or not they account for feature complementarity in the selection process. Brown et al. (2012) however show that both these approaches can be subsumed in a more general, unifying theoretical framework known as conditional likelihood maximization.

3.1 Redundancy-based methods

Most feature selection algorithms in the 1990s and early 2000 focus on relevance and redundancy to obtain the optimal subset. Notable amongst them are (1) mutual information based feature selection (MIFS) (Battiti 1994), (2) correlation based feature selection (CFS) (Hall 2000), (3) minimum redundancy maximum relevance (mRMR) (Peng et al. 2005), (4) fast correlation based filter (FCBF) (Yu and Liu 2003), (5) ReliefF (Kononenko 1994), and (6) conditional mutual information maximization (CMIM) (Fleuret 2004; Wang and Lochovsky 2004). With some variation, their main goal is to maximize relevance and minimize redundancy. Of these methods, MIFS, FCBF, ReliefF, and CMIM are potentially feature ranking algorithms. They rank the features based on certain information maximization criterion (Duch 2006) and select the top k features, where k is decided a priori based on expert knowledge or technical considerations.

MIFS is a sequential feature selection algorithm, in which a candidate feature F_i is selected that maximizes the conditional mutual information $I(F_i; Y|F_S)$. Battiti (1994) approximates this MI by $I(F_i; Y|F_S) = I(F_i; Y) - \beta \sum_{F_j \in F_S} I(F_i; F_j)$, where, F_S is an already selected feature subset, and $\beta \in [0, 1]$ is a user-defined parameter that controls the redundancy. For $\beta = 0$, it reduces to a ranking algorithm. Battiti (1994) finds $\beta \in [0.5, 1]$ is appropriate for many classification tasks. Kwak and Choi (2002) show that when $\beta = 1$, MIFS method penalizes redundancy too strongly, and for this reason does not work well for non-linear dependence.

CMIM implements an idea similar to MIFS, but differs in the way in which $I(F_i; Y|F_S)$ is estimated. CMIM selects the candidate feature F_i that maximizes $\min_{F_j \in F_S} \mathrm{I}(F_i; Y|F_j)$. Both MIFS and CMIM are incremental forward search methods, and they suffer from initial selection bias (Zhang and Zhang 2012). For example, if $\{F_1, F_2\}$ is the selected subset and $\{F_3, F_4\}$ is the candidate subset, the CMIM selects F_3 if $I(F_3; Y|\{F_1, F_2\}) > I(F_4; Y|\{F_1, F_2\})$ and the new optimal subset becomes $\{F_1, F_2, F_3\}$. The incremental search only evaluates the redundancy between the candidate feature F_3 and $\{F_1, F_2\}$ i.e. $I(F_3; Y|\{F_1, F_2\})$ and never considers the redundancy between F_1 and $\{F_2, F_3\}$ i.e. $I(F_1; Y|\{F_2, F_3\})$.

CFS and mRMR are both subset selection algorithms, which evaluate a subset of features using an implicit cost function that simultaneously maximizes relevance and minimizes redundancy. CFS evaluates a subset of features based on pairwise correlation measures, in which correlation is used as a generic measure of dependence. CFS uses the following heuristic to evaluate a subset of features: $merit(S) = \frac{k\,\bar{r}_{cf}}{\sqrt{k + k\,(k-1)\,\bar{r}_{ff}}}$, where k denotes the subset size, $\bar{r}_{cf}$ denotes the average feature-class correlation, and $\bar{r}_{ff}$ denotes the average feature-feature correlation of features in the subset. The feature-feature correlation is used as a measure of redundancy, and feature-class correlation is used as a measure of relevance. The goal of CFS is to find a subset of independent features that are uncorrelated and predictive of the class. CFS ignores feature complementarity, and cannot identify strongly interacting features such as in parity problem (Hall and Holmes 2003).

mRMR is very similar to CFS in principle, however, instead of correlation measures, mRMR uses mutual information $I(F_i; Y)$ as a measure of relevance, and $I(F_i; F_j)$ as a

measure of redundancy. mRMR uses the following heuristic to evaluate a subset of features: $score(S) = \frac{\sum_{i \in S} I(F_i;Y)}{k} - \frac{\sum_{i,j \in S} I(F_i;F_j)}{k^2}$. mRMR method suffers from limitations similar to CFS. Gao et al. (2016) show that the approximations made by the information-theoretic methods, such as mRMR and CMIM, are based on unrealistic assumptions and they introduce a novel set of assumptions based on variational distributions and derive novel algorithms with competitive performance.

FCBF follows a 2-step process. In step 1, it ranks all features based on symmetric uncertainty between each feature and the class variable, i.e., $SU(F_i, Y)$, and selects the relevant features that exceed a given threshold value δ. In step 2, it finds the optimal subset by eliminating redundant features from the relevant features selected in step (i), using an approximate Markov blanket criterion. In essence, it decouples the relevance and redundancy analysis, and circumvents the concurrent subset search and subset evaluation process. Unlike CFS and mRMR, FCBF is computationally fast, simple, and fairly easy to implement due to the sequential 2-step process. However, this method fails to capture situations where feature dependencies appear only conditionally on the class variable (Fleuret 2004). Zhang and Zhang (2012) state that FCBF suffers from instability as the naive heuristics FCBF may be unsuitable in many situations. One of the drawbacks of FCBF is that it rules out the possibility of an irrelevant feature becoming relevant due to interaction with other features (Guyon and Elisseeff 2003). CMIM, which simultaneously evaluates relevance and redundancy at every iteration, overcomes this limitation.

Relief (Kira and Rendell 1992), and its multi-class version ReliefF (Kononenko 1994), are instance-based feature ranking algorithms that rank each feature based on its similarity with k nearest neighbors from the same and opposite classes, selected randomly from the dataset. The underlying principle is that a useful feature should have the same value for instances from the same class and different values for instances from a different class. In this method, m instances are randomly selected from the training data and for each of these m instances, n nearest neighbors are chosen from the same and the opposite class. Values of features of the nearest neighbors are compared with the sample instance and the scores for each feature is updated. A feature has higher weights if it has the same value with instances from the same class and different values to others. In Relief, the score or weight of each feature is measured by the Euclidean distance between the sampled instance and the nearest neighbor, which reflects its ability to discriminate between different classes.

The consistency-based method (Almuallim and Dietterich 1991; Liu and Setiono 1996; Dash and Liu 2003) is another approach, which uses consistency measure as the performance metric. A feature subset is inconsistent if there exist at least two instances with the same feature values but with different class labels. The inconsistency rate of a dataset is the number of inconsistent instances divided by the total number of instances in it. This approach aims to find a subset, whose size is minimal and inconsistency rate is equal to that of the original dataset. Liu and Setiono (1996) propose the following heuristic: $Consistency(S) = 1 - \frac{\sum_{i=0}^{m}(|D_i| - |M_i|)}{N}$, where, m is the number of distinct combinations of feature values for subset S, $|D_i|$ is the number of instances of i-th feature value combination, $|M_i|$ is the cardinality of the majority class of i-th feature value combination, and N is the number of total instances in the dataset.

Markov Blanket (MB) filter (Koller and Sahami 1996) provides another useful technique for variable selection. MB filter works on the principle of conditional independence and excludes a feature only if the MB of the feature is within the set of remaining features. Though the MB framework based on information theory is theoretically optimal in eliminating irrelevant and redundant feature, it is computationally intractable. Incremental association

Markov blanket (IAMB) (Tsamardinos et al. 2003) and Fast-IAMB (Yaramakala and Margaritis 2005) are two MB based algorithms that use conditional mutual information as the metric for conditional independence test. They address the drawback of CMIM by performing redundancy checks during 'growing' (forward pass) and 'shrinkage' phase (backward pass).

3.2 Complementarity-based methods

The literature on complementarity-based feature selection that simultaneously optimize redundancy and complementarity are relatively few, despite earliest research on feature interaction dating back to McGill (1954) and subsequently advanced by Yeung (1991), Jakulin and Bratko (2003), Jakulin and Bratko (2004) and Guyon and Elisseeff (2003). The feature selection methods that consider feature complementarity include double input symmetrical relevance (DISR) (Meyer et al. 2008), redundancy complementariness dispersion based feature selection (RCDFS) (Chen et al. 2015), INTERACT (Zhao and Liu 2007), interaction weight based feature selection (IWFS) (Zeng et al. 2015), and maximum relevance maximum complementary (MRMC) (Chernbumroong et al. 2015), joint mutual information (JMI) (Yang and Moody 1999; Meyer et al. 2008), and min-Interaction Max-Relevancy (mIMR) (Bontempi and Meyer 2010).

The goal of DISR is to find the best subset of a given size d, where d is assumed to be known a priori. It considers complementarity 'implicitly' (Bontempi and Meyer 2010), which means they consider the net effect of redundancy and complementarity in the search process. As a result, DISR does not distinguish between two subsets S_1 and S_2, where S_1 has information gain$= 0.9$ and information loss$= 0.1$, and S_2 has information gain $= 0.8$ and information loss $= 0$. In other words, information loss and information gain are treated equally. DISR works on the principle of k-average sub-subset information criterion, which is shown to be a good approximation of the information of a set of features. They show that the mutual information between a subset $\mathbf{F}_S$ of d features, and the class variable Y is lower bounded by the average information of its subsets. Using notations, $\frac{1}{k!\binom{d}{k}} \sum_{V \subseteq S:|V|=k} I(\mathbf{F}_V; Y) \leq I(\mathbf{F}_S; Y)$. k is considered to be size of the sub-subset such that there is no complementarities of order greater than k. Using $k = 2$, DISR recursively decomposes each bigger subset ($d > 2$) into subsets containing 2 features F_i and F_j ($i \neq j$), and chooses a subset $\mathbf{F}_S$ such that $\mathbf{F}_S = \arg\max_S \sum_{i,j,i<j} I(F_i, F_j; Y)/\binom{d}{2}$. An implementation of this heuristic, known as *MASSIVE* is also proposed.

mIMR method presents another variation of DISR in that (1) mIMR first removes all features that have zero mutual information with the class, and (2) it decomposes the multivariate term in DISR into a linear combination of relevance and interaction terms. mIMR considers causal discovery in the selection process; restricts the selection to variables that have both positive relevance and negative interaction. Both DISR and mIMR belong to a framework, known as Joint Mutual Information (JMI) initially proposed by Yang and Moody (1999). JMI provides a sequential feature selection method in which the JMI score of the incoming feature F_k is given by $J_{jmi}(F_k) = \sum_{F_i \in \mathbf{F}_S} I(F_k, F_i; Y)$ where F_S is the already selected subset. This is the information between the targets and the joint random variable (F_k, F_i) defined by pairing the candidate F_k with each feature previously selected.

In RCDFS, Chen et al. (2015) suggest that ignoring higher order feature dependence may lead to false positives (FP) (actually redundant features misidentified as relevant due to pairwise approximation) being selected in the optimal subset, which may impair the selection

of subsequent features. The degree of interference depends on the number of FPs present in the already selected subset and their degree of correlation with the incoming candidate feature. Only when the true positives (TPs) and FPs have opposing influence on the candidate feature, the selection is misguided. For instance, if the candidate feature is redundant to the FPs but complementary to the TPs, then new feature will be discouraged from selection, while it should be ideally selected and vice-versa. They estimate the interaction information (complementarity or redundancy) of the candidate feature with each of already selected features. They propose to measure this noise by standard deviation (dispersion) of these interaction effects and minimize it. The smaller the dispersion, the less influential is the interference effect of false positives.

One limitation of RCDFS is that it assumes that all TPs in the already selected subset will exhibit a similar type of association, i.e., either all are complementary to, or all are redundant with, the candidate feature (see Figure 1 in Chen et al. 2015). This is a strong assumption and need not be necessarily true. In fact, it is more likely that different dispersion patterns could be observed. In such cases, the proposed method will fail to differentiate between the 'good influence' (due to TPs) and 'bad influence' (due to FPs) and therefore will be ineffective in mitigating the interference effect of FPs in the feature selection process.

Zeng et al. (2015) propose a complementarity-based ranking algorithm, IWFS. Their method is based on interaction weight factors, which reflect the information on whether a feature is redundant or complementary. The interaction weight for a feature is updated at each iteration, and a feature is selected if its interaction weight exceeds a given threshold. Another complementarity-based method, INTERACT, uses a feature sorting metric using data consistency. The c-contribution of a feature is estimated based on its inconsistency rate. A feature is removed if its c-contribution is less than a given threshold of c-contribution, otherwise retained. This method is computationally intensive and has worst-case time complexity $O(N^2M)$, where N is the number of instances and M is the number of features. MRMC method presents a neural network based feature selection that uses relevance and complementary score. Relevance and complementary scores are estimated based on how a feature influences or complements the networks.

Brown et al. (2012) propose a space of potential criterion that encompasses several redundancy and complementarity-based methods. They propose that the worth of a candidate feature F_k given already selected subset $\mathbf{F}_S$ can be represented as $J(F_k) = I(F_k; Y) - \beta \sum_{F_i \in \mathbf{F}_S} I(F_k; F_i) + \gamma \sum_{F_i \in \mathbf{F}_S} I(F_k; F_i|Y)$. Different values of β and γ lead to different feature selection methods. For example, $\gamma = 0$ leads to MIFS, $\beta = \gamma = \frac{1}{|S|}$ lead to JMI, $\gamma = 0$, and $\beta = \frac{1}{|S|}$ lead to mRMR.

4 Motivation and the proposed heuristic

In this section, we first outline the motivation behind using redundancy and complementarity 'explicitly' in the search process and the use of an implicit utility function approach. Then, we propose a heuristic, called self-adaptive feature evaluation (**SAFE**). SAFE is motivated by the implicit utility function approach in multi-objective optimization. Implicit utility function approach belongs to interactive methods of optimization (Roy 1971; Rosenthal 1985), which combines the search process with the decision maker's relative preference over multiple objectives. In interactive method, decision making and optimization occur simultaneously.

Table 1 Golf dataset

Outlook (F1)	Temperature (F2)	Humidity (F3)	Windy (F4)	Play golf (Y)
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

4.1 Implicit versus explicit measurement of complementarity

Combining negative (complementarity) and positive (redundancy) interaction information may produce inconsistent results when the goal is to find an optimal feature subset. In this section, we demonstrate this using the 'Golf' dataset presented in Table 1. The dataset has four features $\{F_1, \ldots, F_4\}$. The information content $I(\mathbf{F}_S; Y)$ of each possible subset $\mathbf{F}_S$ is estimated (1) first, using Eq. (1), and (2) then using the aggregate approach. In the aggregate approach, to compute $I(\mathbf{F}_S; Y)$, we take the average of mutual information of each sub-subsets of two features in $\mathbf{F}_S$ with the class variable, i.e., $I(F_i, F_j; Y)$. For example, $I(F_1, F_2, F_3; Y)$ is approximated by the average of $I(F_1, F_2; Y)$, $I(F_1, F_3; Y)$, and $I(F_2, F_3; Y)$. Table 2 presents the results. Mutual information is computed using *infotheo* package in R and empirical entropy estimator.

Our goal is to find the optimal subset regardless of the subset size. Clearly, in this example, the $\{F_1, F_2, F_3, F_4\}$ is the optimal subset that has maximum information about the class variable. However, using aggregate approach, $\{F_1, F_3, F_4\}$ is the best subset. Moreover, in the aggregate approach, one would assign a higher rank to the subset $\{F_1, F_2, F_3\}$ as compared to $\{F_1, F_2, F_4\}$, though the latter subset has higher information content than the former.

4.2 A new adaptive heuristic

We first introduce the following notations for our heuristic and then define the adaptive cost function.

Subset Relevance Given a subset S, subset relevance, denoted by A_S, is defined by summation of all pairwise mutual information between each feature and the class variable, i.e., $A_S = \sum_{i \in S} I(F_i; Y)$. A_S measures the predictive ability of each individual feature acting alone.

Table 2 Mutual information between a subset F_S and the class Y

No.	Subset (S)	$I(F_S; Y)$	Aggregate approach
1	$\{F_1\}$	0.1710	
2	$\{F_2\}$	0.0203	
3	$\{F_3\}$	0.1052	
4	$\{F_4\}$	0.0334	
5	$\{F_1, F_2\}$	0.3173	
6	$\{F_1, F_3\}$	0.4163	
7	$\{F_1, F_4\}$	0.4163	
8	$\{F_2, F_3\}$	0.1567	
9	$\{F_2, F_4\}$	0.1435	
10	$\{F_3, F_4\}$	0.1809	
11	$\{F_1, F_2, F_3\}$	0.5938	0.2967
12	$\{F_1, F_2, F_4\}$	0.6526	0.2924
13	$\{F_1, F_3, F_4\}$	0.7040	0.3378
14	$\{F_2, F_3, F_4\}$	0.3223	0.1604
15	$\{F_1, F_2, F_3, F_4\}$	0.9713	0.2719

Subset Redundancy Given a subset S, subset redundancy, denoted by R_S, is defined by the summation of all positive 3-way interactions in the subset, i.e., $R_S = \sum_{i,j \in S, i<j} (I(F_i; F_j) - I(F_i; F_j|Y)) \; \forall \; (i, j)$ such that $I(F_i; F_j) > I(F_i; F_j|Y)$. R_S measures information loss due to feature redundancy.

Subset Complementarity Given a subset S, subset complementarity, denoted by C_S, is defined by the absolute value of the sum of all negative 3-way interactions in the subset, i.e., $C_S = \sum_{i,j \in S, i<j} (I(F_i; F_j|Y) - I(F_i; F_j)) \forall \; (i, j)$ such that $I(F_i; F_j) < I(F_i; F_j|Y)$. C_S measures information gain due to feature complementarity.

Subset Dependence Given a subset S, subset dependence, denoted by D_S, is defined by the summation of mutual information between each pair of features, i.e., $D_S = \sum_{i,j \in S, i<j} I(F_i; F_j)$. We use D_S as a measure of unconditional feature redundancy in our heuristic. This is the same as the unconditional mutual information between features described as redundancy in the literature (Battiti 1994; Hall 2000; Peng et al. 2005). We call this subset dependence to distinguish this from the information loss due to redundancy (R_S), which is measured by the difference between conditional and unconditional mutual information. Below, we present the proposed heuristic.

$$Score(S) = \frac{A_S + \gamma \; C_S^{\frac{\beta}{|S|}}}{\sqrt{|S| + \beta \; D_S}} \tag{3}$$

where, A_S, C_S, and D_S are subset relevance, subset complementarity, and subset dependence, respectively, and $|S|$ denotes the subset size. β and γ are hyper-parameters defined as follows: $\alpha = \frac{R_S}{R_S + C_S}$, $\beta = (1 + \alpha)$, $\xi = \frac{C_S}{C_S + A_S}$ and $\gamma = (1 - \xi)$. R_S measures subset redundancy as defined in Definition 4.2. As mentioned above, the heuristic characterizes an adaptive objective function, which evolves depending on the level of feature interactions. We model this adaptation using two hyper-parameters β and γ. The values of these parameters are computed by the heuristic during the search process based on the relative values of relevance, redundancy and complementarity.

The ratio $\alpha \in [0, 1]$ measures the percentage of redundancy in the subset, which determines whether the subset is predominantly redundant or complementary. If $\alpha = 1$, all features are pairwise redundant, we call the subset predominantly redundant. At the other extreme, if $\alpha = 0$ all features are pairwise complementary, we call the subset predominantly complementary. We consider $\alpha = 0/0 = 0$ for a fully independent subset of features which is, however, rarely the case. The hyper-parameter β controls the trade-off between redundancy and complementarity based on the value of α, which is a function of subset characteristics. We consider β as a linear function of α such that the penalty for unconditional dependency increases linearly to twice its value when the subset is fully redundant. This resembles the heuristic of CFS (Hall 2000) when $\alpha = 1$. The $|S|$ in the denominator allows the heuristic to favor smaller subsets, while the square root in the denominator allows the penalty term to vary exponentially with increasing subset size and feature dependency.

The proposed heuristic adaptively modifies the trade-off rule as the search process for the optimal subset continues. We explain how such an adaptive criterion works. As α increases, the subset becomes more redundant, the value of subset complementarity (C_S) decreases ($C_S = 0$ when $\alpha = 1$) leaving us with little opportunity to use complementarity very effectively for feature selection process. In other words, the value of subset complementarity C_S is not sufficiently large to be able to differentiate between the two subsets. At best, we expect to extract a set of features that are less redundant or nearly independent. As a result, $\beta \in [1, 2]$ increasingly penalizes the subset redundancy term D_S in the denominator and rewards subset complementarity C_S in the numerator as α increases from 0 to 1.

In contrast, as α decreases, the subset becomes predominantly complementary leading to an increase in the value C_S. As the magnitude of C_S gets sufficiently large, complementarity plays the key role in differentiating between two subsets compared to subset dependence D_S in the denominator. This, however, causes a bias towards larger subset as the complementarity gain increases monotonically with the size of the subset. We observe that C_S increases exponentially with the logarithm of the subset size as α decreases. In Fig. 6, we demonstrate this for three different real data sets having different degrees of redundancy (α). In order to

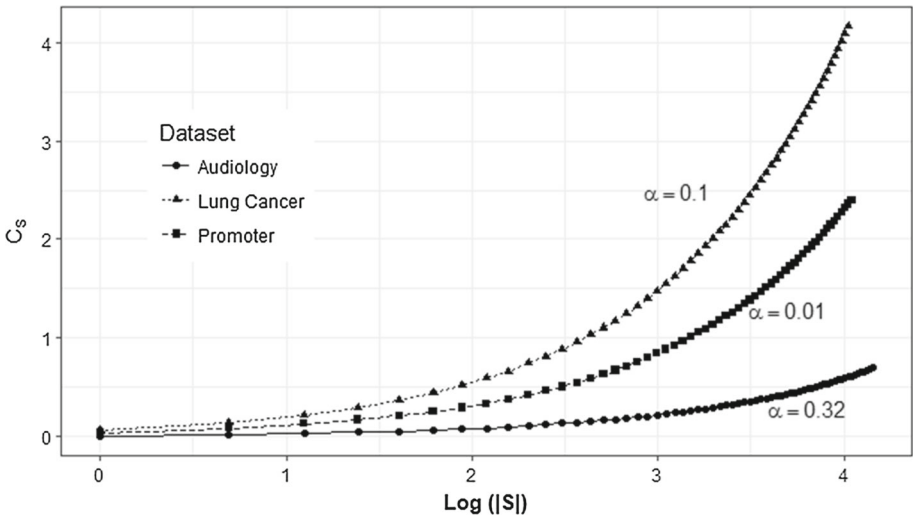

Fig. 6 Variation of subset complementarity with subset size

control this bias, we raise C_S to the exponent $\frac{1}{|S|}$. Moreover, given the way we formalize α, it is possible to have two different subsets, both having $\alpha = 0$ but with different degrees of information gain C_S. This is because the information gain C_S of a subset is affected by both the number of complementary pair of features in the subset as well as how complementary each pair of features is, which is an intrinsic property of each feature. Hence it is possible that a larger subset with weak complementarity produces the same amount of information gain C_S as another smaller subset with highly complementary features. This is evident from Fig. 6, which shows that subset complementarity gains faster for dataset 'Lung Cancer' compared to data set 'Promoter,' despite the fact that the latter has lower α, and both have an identical number of features. The exponent $\frac{1}{|S|}$ also takes care of such issues.

Next, we examine scenarios when $C_S > A_S$. This is the case when a subset learns more information from interaction with other features compared to their individual predictive power. Notice that the source of class information, whether C_S or A_S, is indistinguishable to the heuristic as they are linearly additive in the numerator. As a result, this produces an undue bias towards larger subset size when $C_S > A_S$. To control this bias, we introduce the hyper-parameter $\gamma \in [0, 1]$ that maintains the balance between relevance and complementarity information gain. It controls the subset size by reducing the contribution from C_S when $C_S > A_S$.

4.3 Sensitivity analysis

In Figs. 7 and 8, we show how the proposed heuristic score varies with degree of redundancy α, and the subset size under different relative magnitude of interaction information ($R_S + C_S$), and subset relevance A_S. Figure 7 depicts a situation in which features are individually highly relevant, but less interactive ($C_S < A_S$), whereas in Fig. 8, features are individually less relevant, but as a subset become extremely relevant due to high degree of feature interaction ($C_S > A_S$). In either scenario, the score decreases with increasing subset size, and with increasing degree of redundancy. For a given subset size, the score is generally lower when

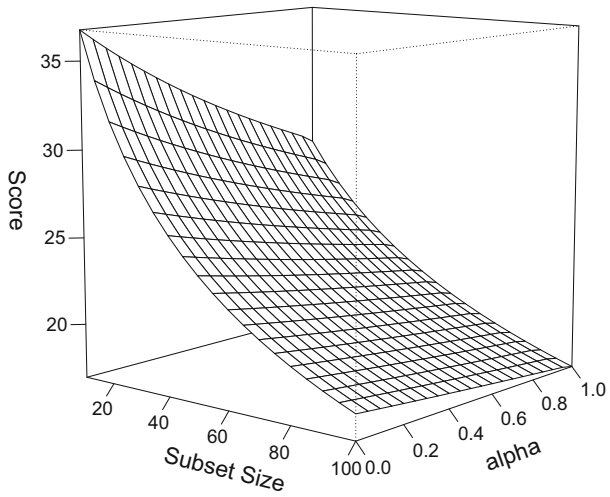

Fig. 7 Heuristic score variation with degree of redundancy α and subset size, given $A_S = 200$, $D_S = 20$, and $R_S + C_S = 20$

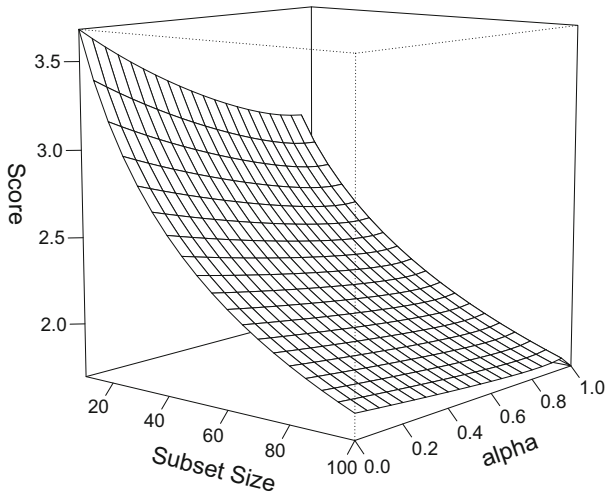

Fig. 8 Heuristic score variation with degree of redundancy α and subset size, given $A_S = 20$, $D_S = 20$, and $R_S + C_S = 200$

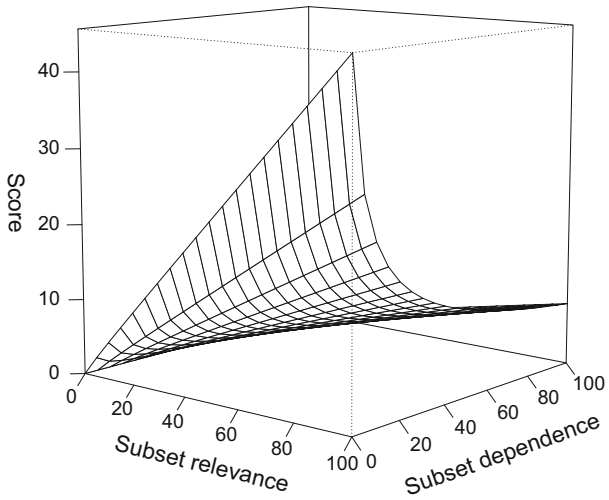

Fig. 9 Heuristic score variation with subset dependence and subset relevance, given $\alpha = 0.5$, $|S| = 5$, and $R_S + C_S = 20$

$C_S > A_S$ compared to when $C_S < A_S$, showing the heuristic is effective in controlling the subset size when the subset is predominantly complementary. We also observe that the heuristic is very sensitive to redundancy when the features are highly relevant. In other words, redundancy hurts much more when features are highly relevant. This is evident from the fact that the score reduces at a much faster rate with increasing redundancy when A_S is very high compared to C_S as in Fig. 7.

In Figs. 9 and 10, we show how the proposed heuristic score varies with subset relevance and subset dependence for two different subset sizes. The score increases linearly with the subset relevance, and decreases non-linearly with the subset dependence, as we would expect

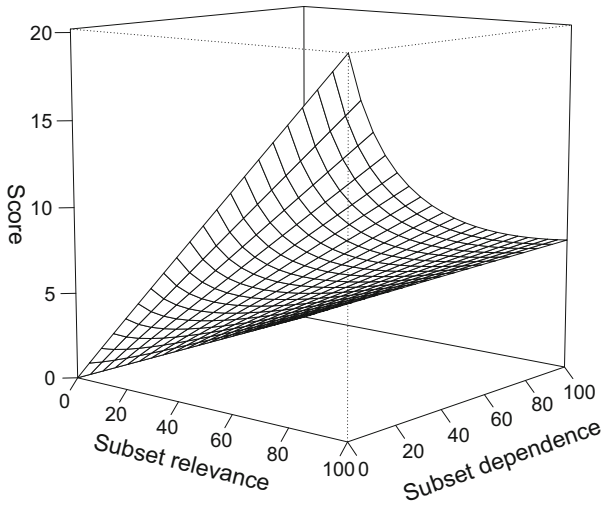

Fig. 10 Heuristic score variation with subset dependence and subset relevance, given $\alpha = 0.5$, $|S| = 25$, and $R_S + C_S = 20$

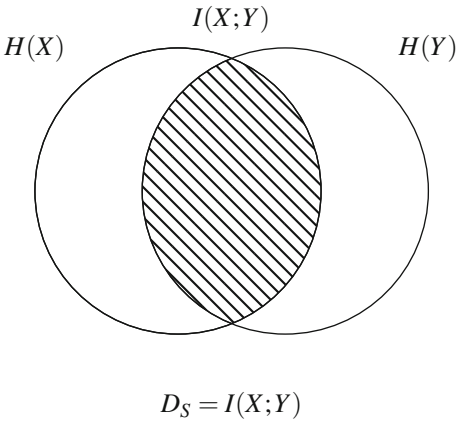

Fig. 11 Fewer features in the subset for given subset dependence D_S

from Eq. 3. The score is higher for a smaller subset compared to a bigger subset under nearly identical conditions. However, for a given subset relevance, the score decreases at a much faster rate with increasing subset dependence when there are fewer number of features in the subset (as in Fig. 9). This phenomenon can be explained with the help of Figs. 11 and 12. For a given subset dependence D_S, fewer features would mean higher degree of association (overlap) between features. In other words, features share more common information, and are therefore more redundant compared to when there are higher number of features in the subset. Hence, our heuristic not only encourages parsimony, but is also sensitive to the change in feature dependence as the subset size changes. The above discussion demonstrates the adaptive nature of the heuristic under different conditions of relevance and redundancy, which is the motivation behind our heuristic.

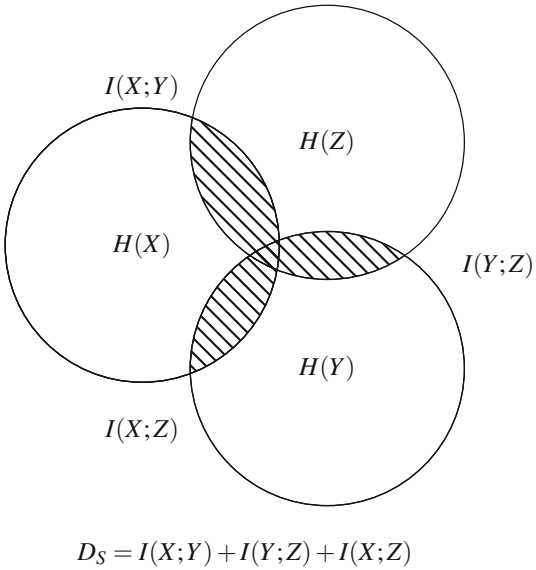

Fig. 12 More features in the subset for given subset dependence D_S

4.4 Limitations

One limitation of this heuristic, as evident from Figs. 9 and 10, is that it assigns a zero score to a subset when the subset relevance is zero, i.e., $A_S = 0$. Since feature relevance is a non-negative measure, this implies a situation in which every feature in the subset is individually irrelevant to the target concept. Thus, our heuristic does not select a subset when none of the features in the subset carry any useful class information by themselves. This, however, ignores the possibility that they can become relevant due to interaction. However, we have not encountered datasets where this is the case.

Another limitation of our heuristic is that it considers up to 3-way feature interaction (interaction between a pair of features and the class variable) and ignores the higher-order corrective terms. This pairwise approximation is necessary for computational tractability and is considered in many MI-based feature selection methods (Brown 2009). Despite this limitation, Kojadinovic (2005) shows that Eq. 1 produces reasonably good estimates of mutual information for all practical purposes. The higher order corrective terms become significant only when there exists a very high degree of dependence amongst a large number of features. It may be noted that our definition of subset complementarity and subset redundancy, as given in Sect. 4.2, can be extended to include higher-order interaction terms without any difficulty. With more precise estimates of mutual information becoming available, further work will address the merit of using higher order correction terms in our proposed approach.

5 Algorithm

In this section, we present the algorithm for our proposed heuristic SAFE.

Step 1 Assume we start with a training sample $D(\mathbf{F}, Y)$ with full feature set $\mathbf{F} = \{F_1, \ldots, F_n\}$ and class variable Y. Using a search strategy, we choose a candidate subset of features $\mathbf{S} \subset \mathbf{F}$.

Step 2 Using the training data, we compute the mutual information between each pair of features $I(F_i; F_j)$, and between each feature and the class variable $I(F_i; Y)$. We eliminate all constant valued features from $\mathbf{S}$, for which $I(F_i; Y) = 0$.
Step 3 For each pair of features in $\mathbf{S}$, we compute the conditional mutual information given the class variable, $I(F_i; F_j|Y)$.
Step 4 We transform all $I(F_i; F_j)$ and $I(F_i; F_j|Y)$ to their symmetric uncertainty form to maintain a scale between [0,1].
Step 5 For each pair of features, we compute the interaction gain or loss, i.e., $I(F_i; F_j; Y) = I(F_i; F_j) - I(F_i; F_j|Y)$ using the symmetric uncertainty measures.
Step 6 We compute A_S, D_S, C_S, R_S, β, and γ using information from Steps 4 & 5.
Step 7 Using information from Step 6, the heuristic determines a $Score(\mathbf{S})$ for subset $\mathbf{S}$. The search continues, and a subset $\mathbf{S}_{opt}$ is chosen that maximizes this score.

The pseudo-algorithm of the proposed heuristic is presented in Algorithm 1.

Algorithm 1 Self Adapting Feature Evaluation (SAFE)

Input: A training sample $D(F, Y)$ with full feature set $F = \{F_1, \ldots, F_n\}$ and class variable Y
Output: The selected feature subset S_{opt}
Initialize $Score_{opt}$, $S_{opt} \leftarrow \emptyset$;
repeat
  Generate a subset $S \subset F$ using a search strategy;
  for i,j = 1 to n; j> i **do**
    Calculate $SU(F_i, Y)$, $SU(F_i, F_j)$, and $SU(F_i, F_j, Y)$;
    Remove all features F_i from S if $SU(F_i, Y) = 0$;
  end for
  Calculate A_S, D_S, C_S, β, γ;
  Calculate $Score(S)$;
  if $Score(S) > Score_{opt}$ **then** $Score_{opt} \leftarrow Score(S)$, $S_{opt} \leftarrow S$;
  else no change in $Score_{opt}$, S_{opt};
  end if
until $Score_{opt}$ does not improve based on stopping rule
Return S_{opt};
End

5.1 Time complexity

The proposed heuristic provides a subset evaluation criterion, which can be used as the heuristic function for determining the score of any subset in the search process. Generating candidate subsets for evaluation is generally a heuristic search process, as searching 2^n subsets is computationally intractable for large n. As a result, different search strategies, such as sequential, random, and complete are adopted. In this paper, we use the best-first search (BFS) (Rich and Knight 1991) to select candidate subsets using the heuristic as the evaluation function.

BFS is a sequential search that expands on the most promising node according to some specified rule. Unlike depth-first or breadth-first method, which selects a feature blindly, BFS carries out informed search and expands the tree by spitting on the feature that maximizes the heuristic score, and allows backtracking during the search process. BFS moves through the search space by making small changes to the current subset, and is able to backtrack to previous subset if that is more promising than the path being searched. Though BFS is

exhaustive in pure form, by using a suitable stopping criterion, the probability of searching the entire feature space can be considerably minimized.

To evaluate a subset of k features, we need to estimate $k(k-1)/2$ mutual informations between each pair of features, and k mutual information between each feature and the class variable. Hence, the time complexity of this operation is $O(k^2)$. To compute the interaction information, we need $k(k-1)/2$ linear operations (substraction). Hence the worst time complexity of the heuristic is $O(n^2)$, when all features are selected. However, this case is rare. Since best-first is forward sequential search method, there is no need to pre-compute $n \times n$ matrix of mutual information pairs in advance. As the search progresses, the computation is done progressively, requiring only incremental computations at each iteration. Using a suitable criterion (maximum number of backtracks), we can restrict the time complexity of BFS. For all practical data sets, the best-first search converges to a solution quickly. Despite that, the computational speed of the heuristic slows down as the number of input features become very large requiring more efficient computation of mutual information. Other search methods such as forward search or branch and bound (Narendra and Fukunaga 1977) method can also be used.

6 Experiments on artificial datasets

In this section, we evaluate the proposed heuristic using artificial data sets. In our experiments, we compare our method with 11 existing feature selection methods: CFS (Hall 2000), ConsFS (Dash and Liu 2003), mRMR (Peng et al. 2005), FCBF (Yu and Liu 2003), ReliefF (Kononenko 1994), MIFS (Battiti 1994), DISR (Meyer and Bontempi 2006), IWFS (Zeng et al. 2015), mIMR (Bontempi and Meyer 2010), JMI (Yang and Moody 1999), and IAMB (Tsamardinos et al. 2003). For IAMB, 4 different conditional independence tests ("mi","mi-adf","χ^2","χ^2-adf") are considered and the union of each Markov blanket is considered as the feature subset. Experiments using artificial datasets help us to validate how well the heuristic deals with irrelevant, redundant, and complementary features because the salient features and the underlying relationship with the class variable are known in advance. We use two multi-level data sets D_1 and D_2 from Doquire and Verleysen (2013) for our experiment. Each dataset has 1000 randomly selected instances, 4 labels, and 8 classes. For the feature ranking algorithms, such as ReliefF, MIFS, IWFS, DISR, JMI, mIMR, we terminate when $I(\mathbf{F}_S; Y) \approx I(\mathbf{F}; Y)$ estimated using Eq. 1, i.e., when all the relevant features are selected (Zeng et al. 2015). For large datasets, however, this information criteria may be time-intensive. Therefore, we restrict the subset size to a maximum of 50% of the initial number of features when we test our heuristic on real data sets in Sect. 7. For example, Zeng et al. (2015) restrict to a maximum of 30 features since the aim of feature selection is to select a smaller subset from the original features. For subset selection algorithms, such as CFS, mRMR, and SAFE, we use best-first search for subset generation.

6.1 Synthetic datasets

D_1: The data set contains 10 features $\{f_1, \ldots, f_{10}\}$ drawn from a uniform distribution on the [0, 1] interval. 5 supplementary features are constructed as follows: $f_{11} = (f_1 - f_2)/2$, $f_{12} = (f_1 + f_2)/2$, $f_{13} = f_3 + 0.1$, $f_{43} = f_4 - 0.2$, and $f_{15} = 2\, f_5$. The multi-label output $O = [O^1 \ldots O^4]$ is constructed by concatenating the four binary outputs O^1 through O^4 evaluated as follows. This multi-label output O is the class variable Y for the classification

problem, which has 8 different class labels. For example, [1001] represents a class label formed by $O^1 = 1$, $O^2 = 0$, $O^3 = 0$, and $O^4 = 1$.

$$\begin{cases} O^1 = 1 & \text{if} \quad f_1 > f_2 \\ O^2 = 1 & \text{if} \quad f_4 > f_3 \\ O^3 = 1 & \text{if} \quad O^1 + O^2 = 1 \\ O^4 = 1 & \text{if} \quad f_5 > 0.8 \\ O^i = 0 & \text{otherwise} \quad (i = 1, 2, 3, 4) \end{cases} \tag{4}$$

The relevant features are f_{11} (or f_1 and f_2), f_3 (or f_{13}), f_4 (or f_{14}), and f_5 (or f_{15}). Remaining features are irrelevant for the class variable.
D_2: The data set contains 8 features $\{f_1, \ldots, f_8\}$ drawn from a uniform distribution on the [0, 1] interval. The multi-label output $O = [O^1 \ldots O^4]$ is constructed as follows:

$$\begin{cases} O^1 = 1 & \text{if} \quad (f_1 > 0.5 \text{ and } f_2 > 0.5) \text{ or } (f_1 < 0.5 \text{ and } f_2 < 0.5) \\ O^2 = 1 & \text{if} \quad (f_3 > 0.5 \text{ and } f_4 > 0.5) \text{ or } (f_3 < 0.5 \text{ and } f_4 < 0.5) \\ O^3 = 1 & \text{if} \quad (f_1 > 0.5 \text{ and } f_4 > 0.5) \text{ or } (f_1 < 0.5 \text{ and } f_4 < 0.5) \\ O^4 = 1 & \text{if} \quad (f_2 > 0.5 \text{ and } f_3 > 0.5) \text{ or } (f_2 < 0.5 \text{ and } f_3 < 0.5) \\ O^i = 0 & \text{otherwise} \quad (i = 1, 2, 3, 4) \end{cases} \tag{5}$$

The relevant features are f_1 to f_4. Remaining features are irrelevant for the class variable. The dataset D_2 reflects higher level of feature interaction. The features are relevant only if considered in pairs. For example, the features f_1 and f_2 together define the class O^1, neither f_1 nor f_2 alone can do. The same observation applies to other pairs: (f_3, f_4), (f_1, f_4), and (f_2, f_3).

6.2 Data pre-processing

In this section, we discuss two important data pre-processing steps—imputation and discretization. We also discuss the packages used in the computation of mutual information for our experiments.

6.2.1 Imputation

Missing data arise in almost all statistical analyses due to various reasons. For example, missing values could be completely at random, at random, or not at random (Little and Rubin 2014). In such a situation, we can either discard those observations with missing values, or use expectation-maximization algorithm (Dempster et al. 1977) to estimate parameters of a model in the presence of missing data, or use imputation. Imputation (Hastie et al. 1999; Troyanskaya et al. 2001) provides a way to estimate the missing values of features. There are several methods in the literature for imputation, of which we use kNN method of imputation, which is used widely. kNN imputation method (Batista and Monard 2002) imputes missing values of a feature using the most frequent value from k nearest neighbors for discrete variable and using weighted average of k nearest neighbors for continuous variable, where weights are based on some distance measure between the instance and its nearest neighbors. As some of the real datasets used in our experiments have missing values, we use kNN imputation with $k = 5$ for imputing the missing values. Using higher values of k presents a trade-off between accuracy of imputed values and computation time.

6.2.2 Discretization

Computation of mutual information of continuous features requires the continuous features to be discretized. Discretization refers to the process of partitioning continuous features into some discrete intervals or nominal values. However, there is always some discretization error or information loss, which needs to be minimized. Dougherty et al. (1995) and Kotsiantis and Kanellopoulos (2006) present a survey of various discretization methods present in the literature. In our experiment, we discretize the continuous features into nominal ones using minimum description length (MDL) method (Fayyad and Irani 1993). The MDL principle states the best hypothesis is the one with minimum description length. While partitioning a continuous variable into smaller discrete intervals reduces the value of entropy function, too fine grained partition increases the risk of over-fitting. MDL principle enables us to balance between the number of discrete intervals and the information gain. Fayyad and Irani (1993) use mutual information to recursively define the best bins or intervals coupled with MDL criterion (Rissanen 1986). We use this method to discretize continuous features in all our experiments.

6.2.3 Estimation of mutual information

For all experiments, mutual information is computed using *infotheo* package in R and empirical entropy estimator. The experiments are carried out using a computer with Windows 7, i5 processor, 2.9 GHz, and statistical package R (R Core Team 2013).

6.3 Experimental results

The results of the experiment on synthetic dataset D_1 is given in Table 3. Except IAMB, all feature selection methods are able to select the relevant features. Five out of 12 methods including SAFE are able to select an optimal subset. mRMR selects the maximum number of features including 5 irrelevant features and 1 redundant feature, and IAMB selects only

Table 3 Results of experiment on artificial dataset D_1

Feature	Subset selected	Irrelevant features	Redundant features
SAFE	$\{f_3, f_4, f_5, f_{11}\}$[a]	–	–
CFS	$\{f_3, f_4, f_5, f_{11}\}$[a]	–	–
ConsFS	$\{f_2, f_3, f_4, f_5, f_{11}\}$	–	f_2
mRMR	$\{f_1, f_2, f_5 - f_{11}, f_{13}, f_{14}\}$	$\{f_6 - f_{10}\}$	f_{11}
FCBF	$\{f_{11}, f_{15}, f_3, f_{14}\}$[a]	–	–
ReliefF	$\{f_{11}, f_5, f_{15}, f_3, f_{13}, f_2, f_1, f_{14}\}$	–	$\{f_1, f_2, f_{13}, f_{15}\}$
MIFS ($\beta = 0.5$)	$\{f_{11}, f_5, f_3, f_4\}$[a]	–	
DISR	$\{f_{11}, f_5, f_{15}, f_8, f_3, f_4\}$	$\{f_8\}$	$\{f_{15}\}$
IWFS	$\{f_{11}, f_5, f_3, f_4\}$[a]	–	–
JMI	$\{f_{11}, f_3, f_4, f_{13}, f_{14}, f_5\}$	–	$\{f_{13}, f_{14}\}$
IAMB	$\{f_{11}\}$	–	–
mIMR	$\{f_{11}, f_{13}, f_1, f_2, f_{12}, f_3, f_4, f_{14}, f_5\}$	$\{f_{12}\}$	$\{f_1, f_2, f_3, f_{14}\}$

[a]Denotes $anoptimalsubset$

Table 4 Results of experiment on artificial dataset D_2

Feature	Subset selected	Irrelevant features	Unrepresented class labels
SAFE	$\{f_1, f_2, f_3, f_4\}$[a]	–	–
CFS	$\{f_1 - f_8\}$	$\{f_5 - f_8\}$	–
ConsFS	$\{f_1, f_2, f_4, f_5, f_6, f_8\}$	$\{f_5, f_6, f_8\}$	O^2, O^4
mRMR	$\{f_1 - f_8\}$	$\{f_5 - f_8\}$	–
FCBF	$\{f_2\}$	–	O^1, O^2, O^3, O^4
ReliefF	$\{f_2, f_3, f_1, f_4\}$[a]	–	–
MIFS ($\beta = 0.5$)	$\{f_2, f_1, f_5, f_4, f_6, f_7, f_3\}$	$\{f_5, f_6, f_7\}$	–
DISR	$\{f_2, f_4, f_1, f_3\}$[a]	–	–
IWFS	$\{f_2, f_4, f_1, f_3\}$[a]	–	–
JMI	$\{f_4, f_2, f_1, f_3\}$[a]	–	–
IAMB	$\{f_1, f_4, f_7, f_8\}$	$\{f_7, f_8\}$	O^1, O^2, O^4
mIMR	$\{f_2, f_8, f_5, f_3\}$	$\{f_8, f_5\}$	O^1, O^2, O^3

[a]Denotes an optimal subset

1 feature. mIMR, and ReliefF select maximum number of redundant features and mRMR selects maximum number of irrelevant features.

The results of experiment on synthetic dataset D_2 is given in Table 4. ConsFS, FCBF, IAMB and mIMR fail to select all relevant features. As discussed in Sect. 6.1, features are pairwise relevant. In the absence of an interactive feature, some apparently useful features become irrelevant, failing to represent some class labels. Those unrepresented class labels are given in the third column of Table 4. Five out of 12 feature selection methods including SAFE are able to select an optimal subset. CFS, ConsFS, mRMR, MIFS, IAMB, and mIMR fail to remove all irrelevant features. FCBF performs poorly on this dataset. The experimental results show that SAFE can identify the relevant and interactive features effectively, and can also remove the irrelevant and redundant features.

7 Experiments on real datasets

In this section, we describe the experimental set-up, and evaluate the performance of our proposed heuristic using 25 real benchmark datasets.

7.1 Benchmark datasets

To validate the performance of the proposed algorithm, 25 benchmark datasets from UCI Machine Learning Repository are used in our experiment, which are widely used in the literature. Table 5 summarizes general information about these datasets. Note that these datasets greatly vary in the number of features (max = 1558, min = 10), type of variables (real, integer and nominal), number of classes (max = 22, min = 2), sample size (max = 9822, min = 32), and extent of missing values, which can provide comprehensive testing, and robustness checks under different conditions.

Table 5 Datasets description

No.	Dataset	Instances	Features	Class	Missing	Baseline accuracy (%)
1	CMC	1473	10	3	No	43.00
2	Wine	178	13	3	No	40.00
3	Vote	435	16	2	Yes	61.00
4	Primary Tumor	339	17	22	Yes	25.00
5	Lymphography	148	19	4	No	55.00
6	Statlog	2310	19	7	No	14.00
7	Hepatitis	155	19	2	Yes	79.00
8	Credit g	1000	20	2	No	70.00
9	Mushroom	8124	22	2	Yes	52.00
10	Cardio	2126	22	10	No	27.00
11	Thyroid	9172	29	21	Yes	74.00
12	Dermatology	366	34	6	Yes	31.00
13	Ionosphere	351	34	2	No	64.00
14	Soybean-s	47	35	4	No	25.00
15	kr-kp	3196	36	2	No	52.00
16	Anneal	898	39	5	Yes	76.00
17	Lung Cancer	32	56	3	Yes	13.00
18	Promoters	106	57	2	No	50.00
19	Splice	3190	60	3	No	50.00
20	Audiology	226	69	9	Yes	25.00
21	CoIL2000	9822	85	2	No	94.00
22	Musk2	6598	166	2	No	85.00
23	Arrhythmia	452	279	16	Yes	54.00
24	CNAE-9	1080	856	9	No	11.11
25	Internet	3279	1558	2	Yes	86.00

Baseline accuracy denotes the classification accuracy obtained when every instance in the whole dataset is classified in the most frequent class

7.2 Validation classifiers

To test the robustness of our method, we use 6 classifiers, naïve Bayes (NB) (John and Langley 1995), logistic regression (LR) (Cox 1958), regularized discriminant analysis (RDA) (Friedman 1989), support vector machine (SVM) (Cristianini and Shawe-Taylor 2000), k-nearest neighbor (kNN) (Aha et al. 1991), and C4.5 (Quinlan 1986; Breiman et al. 1984). These classifiers are not only popular, but also have distinct learning mechanism and model assumptions. The aim is to test the overall performance of the proposed feature selection heuristic for different classifiers.

7.3 Experimental setup

We split each dataset into a training set (70%) and test set (30%) using stratified random sampling. Since the dataset has unbalanced class distribution, we adopt stratified sampling to ensure that both the training and test set represents each class in proportion to their size in the

overall dataset. For the same reason, we choose balanced average accuracy over classification accuracy to measure the classification error rate. Balanced average accuracy is a measure of classification accuracy appropriate for unbalanced class distribution. For a 2-class problem, the balanced accuracy is the average of specificity and sensitivity. For a multi-class problem, it adopts a 'one-versus-all' approach and estimates the balanced accuracy for each class and finally takes the average. Balanced average accuracy takes care of the over-fitting bias of the classifier that results from unbalanced class distribution.

In the next step, each feature selection method is employed to select a smaller subset of features from the original features using the training data. We train each classifier on the training set based on selected features, learn its parameters, and then estimate its balanced average accuracy on the test set. We repeat this process 100 times using different random splits of the dataset, and report the average result. The accuracies obtained using different random splits are observed to be approximately normally distributed. So, the average accuracy of **SAFE** is compared with other methods using paired t test at 5% significance level and significant wins/ties/losses(W/T/L) are reported. Since we compare our proposed method over multiple datasets, the p values have been adjusted using Benjamini–Hochberg adjustments for multiple testing (Benjamini and Hochberg 1995). We also report the number of features selected by each algorithm, and their computation time. Computation time is used as a proxy for the complexity level of the algorithm.

For feature ranking algorithms, we need a threshold to select the optimal subset from the list of ordered features. For ReliefF, we consider the threshold $\delta = 0.05$, which is common in the literature (Hall 2000; Ruiz et al. 2002; Koprinska 2009). For the remaining ones, we terminate when $I(\mathbf{F}_S; Y) \approx I(\mathbf{F}; Y)$ or a maximum of 50% of the initial number of features is selected, whichever occurs earlier. For ReliefF, we consider $k = 5, m = 250$, and exponential decay in weights based on distance (Robnik-Šikonja and Kononenko 2003). The discretized data are used for both training the model and testing its accuracy on the test set.

7.4 Experimental results

In this section, we present the results of the experiment, compare the accuracy, computation time, number of selected features by each method.

7.4.1 Accuracy comparison

Tables 6, 7, 8, 9, 10 and 11 show the balanced average accuracy of 12 different feature selection methods including SAFE, tested with 6 different classifiers for all 25 data sets, resulting in 1800 combinations. For each data set, the best average accuracy is shown in bold font. $a(b)$ in the superscript denotes that our proposed method is significantly better (worse) than the other method using paired t test at 5% significance level after Benjamini–Hochberg adjustment for multiple comparisons. W/T/L denotes the number of datasets in which the proposed method SAFE wins, ties or loses with respect to the other feature selection methods. A summary of wins/ties/loses (W/T/L) results is given in Table 12. The average value of accuracy for each method over all data sets is also presented in the "Avg." row. The consFS method did not converge for 3 data sets within a threshold time of 30 min; those results are not reported.

The results show that SAFE generally outperforms (in terms of W/T/L) other feature selection methods for different classifier-dataset combinations. In all cases except one (MIFS with kNN classifier in Table 10), the number of wins exceeds the number of losses, which shows that our proposed heuristic is effective under different model assumptions and learning

Table 6 Comparison of balanced average accuracy of algorithms using NB classifier

No.	CFS	ConsFS	mRMR	FCBF	ReliefF	MIFS	DISR	IWFS	mIMR	IAMB	JMI	SAFE
1	63.28^a	63.90	62.77^a	62.27^a	63.63	55.48^a	63.81	63.97	63.44^a	63.26^a	63.76	**64.00**
2	96.60^a	97.25^a	97.82^a	98.05	97.44^a	97.46^a	97.74^a	94.57^a	96.90^a	96.74^a	98.07	**98.14**
3	**96.61**	50.08^a	94.28^a	94.99^a	93.22^a	50.08^a	93.25^a	96.58	93.72^a	94.96^a	93.83^a	**96.61**
4	59.73	**60.69**^{b}	59.79	58.56^a	58.95^a	59.25^a	60.19	60.26	58.17^a	56.46^a	57.99^a	59.91
5	68.66	69.30^b	68.45	66.97^a	67.95	**70.35**^{b}	68.86^b	68.01	69.92^b	68.81	68.38	68.17
6	95.22^a	94.11^a	**96.15**^{b}	95.74^a	94.74^a	94.29^a	94.92^a	89.72^a	93.85^a	77.50^a	94.93^a	95.90
7	82.81^a	80.48^a	84.25	80.16^a	**86.64**^{b}	85.21	85.12	82.83^a	85.84^b	81.14^a	84.04	84.35
8	62.07^a	66.94^b	67.33^b	63.73^a	54.11^a	53.50^a	65.08	66.18^b	**67.93**^{b}	64.72	67.30^b	65.15
9	98.44^b	98.93^b	98.96^b	98.43^b	99.63^b	**99.84**^{b}	99.57^b	98.84^b	89.50^a	98.90^b	99.49^b	97.93
10	82.12^a	82.26^a	74.73^a	**85.52**^{b}	82.54^a	82.47^a	82.59^a	79.42^a	83.99^b	78.99^a	76.91^a	83.05
11	77.85	78.49	53.73^a	76.76^a	78.69^b	78.32	77.51^a	72.87^a	**78.82**^{b}	68.35^a	77.52^a	78.14
12	93.33^a	92.12^a	91.95^a	91.37^a	91.77^a	90.94^a	89.29^a	86.06^a	84.41^a	73.04^a	90.95^a	**94.53**
13	91.55	88.66^a	65.19^a	90.69^a	86.04^a	86.04^a	86.04^a	88.14^a	90.40^a	88.24^a	87.56^a	**91.83**
14	96.70	96.07	96.46	**99.99**^{b}	90.82^a	87.12^a	96.42	97.93^b	93.03^a	93.29^a	63.76	96.52
15	89.18^a	**94.20**^{b}	55.81^a	67.43^a	89.92^a	82.21^a	87.56^a	88.62^a	87.78^a	90.49	87.99^a	90.18
16	91.19^a	91.32^a	81.99^a	92.59^a	93.13	91.75^a	92.27^a	91.76^a	88.88^a	75.50^a	93.05	**93.34**
17	58.50^a	58.58^a	58.56^a	58.38^a	**64.38**^{b}	53.20^a	91.76^a	56.85^a	56.18^a	59.63^a	55.92^a	61.47
18	88.78	84.68^a	86.43^a	86.37^a	87.43^a	88.75	86.15^a	88.90	83.25^a	80.56^a	89.16	**89.34**

Table 6 continued

No.	CFS	ConsFS	mRMR	FCBF	ReliefF	MIFS	DISR	IWFS	mIMR	IAMB	JMI	SAFE
19	95.12^{a}	95.77^{a}	**96.96**^{b}	95.19^{a}	95.57^{a}	95.12^{a}	93.83^{a}	93.56^{a}	95.61^{a}	94.01^{a}	95.83^{a}	96.00
20	65.11	67.64^{b}	76.68^{b}	65.16	65.39	65.13	65.22	65.65	65.26	**76.93**^{b}	66.49^{b}	65.29
21	53.72^{b}	*	50.07^{a}	50.30^{a}	59.69^{b}	51.52	62.99^{b}	62.31^{b}	**63.35**^{b}	50.23^{a}	62.76^{b}	51.06
22	83.33	*	50.00^{a}	72.30^{a}	89.86^{b}	82.24^{a}	84.53^{b}	82.37^{a}	83.92^{b}	72.76^{a}	**90.62**^{b}	83.23
23	56.26^{a}	63.94^{a}	59.36^{a}	63.90^{a}	56.30^{a}	57.38^{a}	62.70^{a}	64.13^{a}	60.46^{a}	58.10^{a}	60.81^{a}	**67.00**
24	**86.91**^{b}	*	79.99	78.58^{a}	86.60^{b}	82.78^{b}	76.18^{a}	79.62	76.31^{a}	68.36^{a}	77.98^{a}	80.37
25	83.77^{b}	64.20^{a}	86.38^{b}	**88.48**^{b}	75.63	71.58^{a}	85.29^{b}	78.10^{b}	85.78^{b}	78.57^{a}	88.40^{b}	75.75
Avg.	80.67	79.10	75.76	79.28	80.40	76.48	80.48	79.89	79.87	76.33	80.92	**81.11**
W/T/L	**12**/9/4	**13**/3/6	**14**/5/6	**19**/2/4	**13**/5/7	**17**/5/3	**14**/6/5	**13**/7/3	**16**/1/8	**20**/3/2	**13**/6/6	

Table 7 Comparison of balanced average accuracy of algorithms using LR classifier

No.	CFS	ConsFS	mRMR	FCBF	ReliefF	MIFS	DISR	IWFS	mIMR	IAMB	JMI	SAFE
1	64.34^{a}	64.77	62.60^{a}	62.55^{a}	55.12^{a}	55.52^{a}	**66.16**^{b}	66.09^{b}	65.82^{b}	63.77^{a}	65.10	64.91
2	96.07^{a}	96.38^{a}	96.92	97.44	96.80	**97.52**^{b}	96.65	94.10^{a}	95.51^{a}	95.91^{a}	96.77	97.02
3	96.61	50.00^{a}	96.10^{a}	95.79^{a}	96.04^{a}	96.16^{a}	95.51^{a}	96.61	96.04^{a}	**96.66**	96.02^{a}	96.61
4	59.82^{b}	**61.09**^{b}	60.53^{b}	58.35	59.84^{b}	58.94	57.58^{a}	59.30^{b}	57.89^{a}	56.46^{a}	57.39^{a}	58.42
5	75.83	72.88^{a}	76.67	71.50^{a}	**79.02**^{b}	76.76	73.39^{a}	73.01^{a}	72.24^{a}	68.90^{a}	71.69^{a}	75.57
6	95.64^{a}	93.99^{a}	96.28	96.32	96.26	**96.81**^{b}	96.59^{b}	90.52^{a}	96.23	77.58^{a}	96.77^{b}	96.14
7	79.90	79.13^{a}	79.34^{a}	78.48^{a}	80.28	79.47^{a}	**82.34**^{b}	80.40	80.14	79.57^{a}	80.58	**81.02**
8	60.80^{a}	65.36^{b}	65.27^{b}	62.61^{a}	**65.60**^{b}	53.50^{a}	63.38^{a}	63.05^{a}	65.32^{b}	63.03^{a}	65.53^{b}	64.33
9	98.44^{a}	99.96^{b}	98.96	98.43^{a}	98.96	**99.99**^{b}	99.37^{b}	99.99^{b}	97.25^{a}	99.35^{b}	99.77^{b}	99.06
10	83.10^{a}	85.16^{b}	74.66^{a}	**86.11**^{b}	83.25^{a}	81.32^{a}	83.37^{a}	81.57^{a}	84.36^{b}	78.88^{a}	78.06^{a}	83.94
11	79.00^{a}	**82.27**	53.96^{a}	78.95^{a}	78.42^{a}	81.22^{a}	77.76^{a}	72.37^{a}	80.33^{a}	67.71^{a}	79.37^{a}	81.86
12	96.14^{b}	91.16^{a}	95.03	95.97^{b}	95.59	**96.31**^{b}	89.83^{a}	86.35^{a}	86.37^{a}	72.87^{a}	92.71^{a}	95.27
13	87.97	86.70^{a}	65.19^{a}	88.79	65.19^{a}	**90.88**^{b}	88.44	88.57	89.54	88.15	89.10	88.80
14	99.21^{a}	96.78^{a}	99.11^{a}	**99.99**	99.68	**99.99**	98.32^{a}	97.93^{a}	**99.99**	92.05^{a}	99.76	99.83
15	89.18	94.34^{b}	55.83^{a}	67.43^{a}	89.95^{a}	82.21^{a}	94.23^{b}	**96.54**^{b}	95.39^{b}	94.88^{b}	95.41^{b}	90.18
16	92.08^{a}	93.84^{b}	81.96^{a}	93.15	94.05^{b}	93.92^{a}	**94.73**^{b}	93.03^{a}	85.55^{a}	75.58^{a}	93.76	93.49
17	59.24	57.55^{a}	58.54	58.50	**63.17**^{b}	59.68	58.64	57.95^{a}	62.59	60.17	62.86^{b}	60.00
18	**88.40**	81.84^{a}	80.87^{a}	84.40^{a}	82.96^{a}	87.18	84.21^{a}	87.28	85.56^{a}	80.00^{a}	87.40	88.31
19	95.44^{a}	95.86^{a}	96.52	95.50^{a}	**96.54**^{b}	95.50^{a}	96.23^{a}	93.67^{a}	96.18^{a}	94.67^{a}	96.21^{a}	96.38

Table 7 continued

No.	CFS	ConsFS	mRMR	FCBF	ReliefF	MIFS	DISR	IWFS	mIMR	IAMB	JMI	SAFE
20	73.21^a	75.48	**78.42**^{b}	72.99^a	69.38^a	70.95^a	70.36^a	73.17^a	71.90^a	78.07^b	74.53^a	75.51
21	50.29	*	50.01^a	50.31	50.01^a	50.22	50.73^b	51.58^b	51.32^b	50.05^a	**51.66**^{b}	50.28
22	85.46^b	*	50.00^a	72.31^a	72.98^a	83.87	85.33^b	82.25^a	89.13^b	72.48^a	**90.66**^{b}	84.25
23	66.27^a	64.69^a	66.44^a	66.30^a	63.07^a	67.13	64.77^a	63.59^a	62.50^a	58.07^a	63.81^a	**67.68**
24	**86.25**^{b}	*	79.99	78.85^a	79.86^a	82.26^b	76.04^a	80.07	76.70^a	68.26^a	77.88^a	80.90
25	81.97^b	64.20^a	83.02^b	87.59^b	76.79	71.65^a	87.40	77.25^b	87.60^b	78.86^b	**88.62**^{b}	75.67
Avg.	81.62	77.45	76.09	79.94	79.55	80.36	81.28	80.22	81.26	76.48	**82.06**	81.82
W/T/L	**12**/8/5	**13**/3/6	**13**/8/4	**14**/8/3	**12**/7/6	**11**/8/6	**13**/3/9	**14**/5/6	**13**/5/7	**18**/3/4	**10**/7/8	

a(*b*) represents the proposed method SAFE is significantly better (worse) than the other method at 5% significance level. Figures in bold indicates the result for the best classifier for that dataset. W/T/L denotes the number of datasets in which the proposed method SAFE wins, ties or loses with respect to the other feature selection method at 5% significance level after Benjamini–Hochberg adjustment for multiple comparisons. We use a computer with Windows 7, i5 processor, and the time-out threshold was kept at 30 mins. *Denotes solution not converged within threshold time

Table 8 Comparison of balanced average accuracy of algorithms using RDA classifier

No.	CFS	ConsFS	mRMR	FCBF	ReliefF	MIFS	DISR	IWFS	mIMR	IAMB	JMI	SAFE
1	63.21^{a}	**64.39**	63.61^{a}	61.56^{a}	53.98^{a}	54.39^{a}	64.16	64.16^{a}	63.99^{a}	62.97^{a}	64.29	64.37
2	96.51^{a}	97.59^{a}	97.80^{a}	**98.20**	97.78^{a}	97.16^{a}	97.62^{a}	94.03^{b}	96.56^{a}	96.68^{a}	98.04	98.18
3	**96.61**	50.00^{a}	94.82^{a}	96.59	94.78^{a}	50.00^{a}	94.61^{a}	96.58	94.71^{a}	96.46^{a}	96.18^{a}	**96.61**
4	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00
5	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	**68.35**^{b}	50.00	50.00
6	95.13^{b}	69.83	**96.34**^{b}	86.54^{b}	**96.34**^{b}	50.00^{a}	50.00^{a}	71.64^{b}	50.00^{a}	74.70^{b}	50.00^{a}	71.66
7	82.37	80.16^{a}	79.87^{a}	81.92	80.68^{a}	82.46	**83.32**	81.94	82.26	81.00^{a}	82.22	82.45
8	63.86^{a}	66.81	67.22	64.95^{a}	67.60^{b}	64.95^{a}	**67.81**^{b}	66.96	67.67^{b}	65.43^{a}	67.41^{b}	66.70
9	98.44^{a}	99.51^{b}	98.55^{a}	98.43^{a}	98.55^{a}	**99.90**^{b}	99.86^{b}	99.19^{b}	95.10^{a}	99.30^{b}	99.37^{b}	98.84
10	81.82^{a}	78.98^{a}	74.47^{a}	84.54^{b}	82.36^{a}	50.00^{a}	84.17^{b}	81.64^{a}	**84.97**^{b}	74.54^{a}	77.28^{a}	83.29
11	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00
12	**95.78**^{b}	92.39^{a}	93.77^{a}	95.56^{b}	92.96^{a}	95.40^{b}	89.95^{a}	86.50^{a}	83.44^{a}	72.60^{a}	91.59^{a}	94.73
13	**88.58**^{b}	88.05	65.19^{a}	87.33	65.19^{a}	50.00^{a}	50.31^{a}	88.18	51.42^{a}	87.62	50.35^{a}	87.60
14	96.36^{a}	96.03^{a}	97.75	**99.99**^{b}	98.91^{b}	99.27^{b}	95.27^{a}	97.93	99.40^{b}	92.05^{a}	99.61^{b}	97.95
15	89.18	94.24^{b}	55.82^{a}	67.43^{a}	91.40^{b}	78.74^{a}	89.34^{a}	**94.69**^{b}	91.06^{b}	94.09^{b}	91.88^{b}	90.18
16	90.40^{a}	**93.87**	82.40^{a}	92.38^{a}	93.70	92.05^{a}	92.10^{a}	92.31^{a}	90.91^{a}	83.75^{a}	93.58	93.83
17	59.89^{a}	61.55	**85.11**^{b}	59.87^{a}	61.86	61.34	61.96	58.56^{a}	67.07^{b}	64.51^{b}	61.84^{a}	63.56
18	86.53^{a}	81.96^{a}	78.93^{a}	83.06^{a}	82.78^{a}	86.28^{a}	77.93^{a}	86.78	79.68^{a}	80.00^{a}	86.90	**87.81**
19	94.57^{a}	95.38^{a}	96.98^{b}	94.68^{a}	**97.03**^{b}	94.68^{a}	95.80	92.81^{a}	95.37^{a}	94.14^{a}	95.80	95.81

Table 8 continued

No.	CFS	ConsFS	mRMR	FCBF	ReliefF	MIFS	DISR	IWFS	mIMR	IAMB	JMI	SAFE
20	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00
21	57.58^{a}	*	51.45^{a}	50.92^{a}	51.97^{a}	55.20^{a}	50.00^{a}	50.00^{a}	50.00^{a}	**64.17**^{b}	50.00^{a}	60.67
22	85.79^{b}	*	52.42^{a}	75.96^{a}	75.48^{a}	84.20	**86.77**^{b}	83.80^{a}	50.00^{a}	80.46^{a}	50.00^{a}	84.53
23	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00
24	**85.59**^{b}	*	78.68	77.26	78.68	82.57^{b}	76.18^{a}	77.47	73.95^{a}	68.50^{a}	77.98^{a}	79.10
25	84.71^{b}	64.20^{a}	86.47^{b}	97.02^{b}	79.38^{b}	73.09^{a}	90.00^{b}	79.55^{b}	**91.83**^{b}	79.28^{b}	90.92^{b}	75.72
Avg.	**77.72**	73.86	73.91	75.93	75.66	70.07	73.89	75.79	72.38	75.22	73.01	76.94
W/T/L	**11**/8/6	9/**11**/2	**13**/8/4	**10**/**10**/5	**11**/8/6	**13**/8/4	**11**/9/5	9/**12**/4	**13**/6/6	**13**/5/7	9/**11**/5	

Table 9 Comparison of balanced average accuracy of algorithms using SVM classifier

No.	CFS	ConsFS	mRMR	FCBF	ReliefF	MIFS	DISR	IWFS	mIMR	IAMB	JMI	SAFE
1	63.99^{a}	64.04^{a}	61.63^{a}	62.61^{a}	55.12^{a}	55.51^{a}	**65.05**^{b}	64.61	64.96^{b}	63.58^{a}	64.66^{b}	64.45
2	97.90^{a}	96.86^{a}	99.36	**99.60**	99.37	99.13^{a}	98.93^{a}	91.92^{a}	97.56^{a}	95.94^{a}	98.23^{a}	99.39
3	**96.61**	50.00^{a}	96.43^{a}	96.57	96.42^{a}	96.33^{a}	96.29^{a}	**96.61**	96.30^{a}	96.48	96.33^{a}	**96.61**
4	58.44^{b}	**59.12**^{b}	58.58^{b}	57.10^{a}	57.89	57.93	57.28^{a}	55.29^{a}	56.99^{a}	56.12^{a}	56.65^{a}	57.53
5	64.43^{a}	63.98^{a}	63.38^{a}	63.79^{a}	66.64^{b}	65.97^{b}	65.77^{b}	63.73^{a}	63.99^{a}	**66.80**^{b}	65.04	64.98
6	94.10^{a}	90.73^{a}	97.04^{a}	95.25^{a}	94.82^{a}	97.64	97.16^{a}	87.27^{a}	96.25^{a}	97.32	97.27	**97.65**
7	77.58^{a}	78.65	80.70	78.54	**81.04**^{b}	79.53	80.96^{b}	80.53	78.36	78.70	80.63	79.61
8	61.41^{a}	62.80	63.82	63.04	**64.11**^{b}	52.48^{a}	63.63	54.27^{a}	63.61	63.29	63.78^{b}	63.22
9	98.44^{a}	99.96^{b}	98.96	98.43^{a}	98.96	**99.99**^{b}	99.98^{b}	99.37^{b}	98.67^{a}	99.37^{b}	99.78^{b}	99.06
10	80.97^{a}	82.50	73.38^{a}	76.91^{a}	76.94^{a}	**85.58**^{b}	81.21^{a}	79.24^{a}	82.29^{a}	78.78^{a}	74.49^{a}	82.52
11	64.36^{a}	64.04^{a}	51.90^{a}	64.21^{a}	63.31^{a}	64.37^{a}	**73.04**^{b}	62.74^{a}	72.90^{b}	65.37^{b}	71.31^{b}	**64.60**
12	**96.23**^{b}	90.05^{a}	94.63^{a}	95.90	94.94^{a}	96.17^{b}	87.51^{a}	80.27^{a}	85.01^{a}	72.47^{a}	89.76^{a}	95.57
13	87.92^{a}	87.10^{a}	65.19^{a}	88.51^{a}	65.19^{a}	90.14^{b}	87.59^{a}	84.52^{a}	**90.96**^{b}	86.41^{a}	89.69	89.12
14	99.96	97.10^{a}	97.87^{a}	**99.99**	98.67^{a}	**99.99**	**99.99**	97.93^{a}	99.89	92.05^{a}	**99.99**	99.96
15	89.18^{a}	94.38^{b}	55.81^{a}	67.43^{a}	89.95^{a}	82.21^{a}	94.75^{b}	93.79^{b}	95.87^{b}	**96.45**^{b}	95.86^{b}	90.18
16	85.06^{a}	91.09^{a}	81.86^{a}	86.62^{a}	**91.72**	90.60^{a}	90.91^{a}	76.88^{a}	83.34^{a}	74.58^{a}	90.53^{a}	91.38
17	54.33^{a}	54.17^{a}	57.62	55.74^{a}	52.02^{a}	52.01^{a}	57.31	53.88^{a}	**60.01**^{b}	56.97^{a}	54.74^{a}	58.00
18	89.37^{a}	80.31^{a}	87.93^{a}	83.31^{a}	90.71	**91.21**	88.84^{a}	85.46^{a}	83.59^{a}	78.68^{a}	91.18	90.66
19	95.28^{a}	95.94^{a}	96.63^{b}	94.02^{a}	**96.67**^{b}	95.37^{a}	96.17^{a}	93.37^{a}	96.22^{a}	94.31^{a}	96.37^{a}	96.49

Table 9 continued

No.	CFS	ConsFS	mRMR	FCBF	ReliefF	MIFS	DISR	IWFS	mIMR	IAMB	JMI	SAFE
20	66.41	**67.15**	65.87^a	66.11^a	66.75	65.61^a	65.50^a	58.05^a	65.00^a	64.85^a	66.64	66.59
21	50.00	*	49.99^a	50.00	**50.02**	49.05^a	50.00	50.00	50.00	50.00	50.00	50.00
22	84.10^a	*	50.00^a	71.90^a	72.75^a	83.77^a	85.84^b	77.56^a	**86.99**^{b}	74.91^a	86.94^b	85.14
23	57.97^a	55.05^a	51.22^a	51.60^a	56.98^a	60.18^a	61.14	50.52^a	58.62^a	57.54^a	55.49^a	**61.67**
24.	**86.18**^{b}	*	77.85^a	76.98^a	77.85^a	82.78^b	76.31^a	75.90^a	76.84^a	68.43^a	78.05^a	80.59
25	83.56^b	64.20^a	84.76^b	88.35^b	75.99	71.72^a	88.57^b	65.81^a	88.49^b	79.10^b	**89.46**^{b}	75.67
Avg.	79.35	76.78	74.50	77.32	77.39	78.61	**80.40**	75.18	79.71	76.34	80.11	80.03
W/T/L	**17**/4/4	**15**/4/3	**17**/5/3	**17**/7/1	**13**/8/4	**14**/5/6	**12**/5/8	**19**/4/2	**14**/4/7	**15**/5/5	**10**/8/7	

a(*b*) represents the proposed method SAFE is significantly better (worse) than the other method at 5% significance level. Figures in bold indicates the result for the best classifier for that dataset. W/T/L denotes the number of datasets in which the proposed method SAFE wins, ties or loses with respect to the other feature selection method at 5% significance level after Benjamini–Hochberg adjustment for multiple comparisons. We use a computer with Windows 7, i5 processor, and the time-out threshold was kept at 30 mins. *Denotes solution not converged within threshold time

Table 10 Comparison of balanced average accuracy of algorithms using kNN classifier

No.	CFS	ConsFS	mRMR	FCBF	ReliefF	MIFS	DISR	IWFS	mIMR	IAMB	JMI	SAFE
1	95.84^{a}	**96.50**	**96.50**	**96.50**	96.44	**96.50**	96.48	96.46	**96.50**	**96.50**	96.49	**96.50**
2	98.32^{a}	98.69	**98.72**	98.68	98.70	98.70	97.53^{a}	98.67^{a}	96.61^{a}	95.08^{a}	97.73^{a}	98.71
3	**96.61**	50.00^{a}	94.58^{a}	94.56^{a}	94.59^{a}	94.11^{a}	94.70^{a}	51.14^{a}	95.10^{a}	95.54^{a}	95.13^{a}	**96.61**
4	79.16^{a}	80.44	80.40	80.50	80.49	**80.52**	80.42	80.51	80.51	56.32^{a}	80.42	80.48
5	65.04^{a}	68.32	68.45	68.53	68.42	**68.59**	66.19^{a}	68.45	63.27^{a}	63.59^{a}	63.65^{a}	68.37
6	96.79^{a}	97.25	97.26	97.25	97.24	97.24	96.84^{a}	96.24^{a}	97.26	**97.62**^{b}	97.26	97.27
7	80.17^{a}	77.37^{a}	81.78	78.61^{a}	82.05	80.59^{a}	83.00	**89.78**^{b}	82.97	80.17^{a}	82.47	82.42
8	60.71^{a}	58.02^{a}	61.83^{a}	62.37	61.99^{a}	53.46^{a}	64.07^{b}	61.13^{a}	**64.42**^{b}	63.37	63.38	62.90
9	99.99	99.99	99.99	99.99	99.99	99.99	99.99	99.99	99.99	99.99	99.99	99.99
10	81.82^{a}	**83.62**	83.61	83.60	83.58	83.57	77.77^{a}	83.60	77.24^{a}	76.95^{a}	74.66^{a}	83.60
11	87.51	87.59	87.36	84.82^{a}	87.54	87.44	70.36^{a}	**87.60**	68.07^{a}	65.30^{a}	67.62^{a}	**87.60**
12	85.73^{a}	78.56^{a}	82.34^{a}	87.87	81.37^{a}	87.07	89.44^{b}	87.96	86.07^{a}	72.47^{a}	**91.99**^{b}	87.33
13	85.58^{a}	83.32^{a}	65.19^{a}	82.02^{a}	65.19^{a}	**90.32**^{b}	76.65^{a}	73.13^{a}	78.14^{a}	85.83^{a}	73.21^{a}	87.44
14	94.33^{a}	96.55	96.44	94.44^{a}	96.66	96.77	**99.99**^{b}	96.44	95.50	90.38^{a}	96.47	96.56
15	99.43^{a}	**99.64**	**99.64**	**99.64**	94.13^{a}	**99.64**	99.46^{a}	**99.64**	94.44^{a}	95.64^{a}	94.56^{a}	**99.64**
16	85.31^{a}	89.69	**89.90**	89.63	89.82	89.79	78.21^{a}	89.77	77.21^{a}	73.10^{a}	79.44^{a}	89.76
17	64.43^{a}	69.24	**70.76**^{b}	68.54	69.00	69.40	65.92^{a}	69.14	68.21	58.55^{a}	65.11^{a}	68.91
18	83.98^{a}	77.75^{a}	71.47^{a}	83.71^{a}	78.21^{a}	81.84^{a}	80.15^{a}	82.40^{a}	79.59^{a}	79.50^{a}	84.50^{a}	**86.53**
19	**92.44**^{b}	90.91^{a}	90.67^{a}	92.34^{b}	90.94^{a}	92.30^{b}	89.07^{a}	90.75^{a}	88.86^{a}	91.15	89.12^{a}	91.17
20	59.39^{a}	61.19^{a}	58.84^{a}	59.80^{a}	59.21^{a}	59.58^{a}	59.90^{a}	63.88^{a}	65.00^{a}	64.85^{a}	**66.64**	66.59

Table 10 continued

No.	CFS	ConsFS	mRMR	FCBF	ReliefF	MIFS	DISR	IWFS	mIMR	IAMB	JMI	SAFE
21	50.96	*	**51.06**	**51.06**	51.00	**51.06**	50.79[a]	**51.06**	50.80[a]	49.99[a]	50.62[a]	**51.06**
22	88.34[a]	*	**89.43**	89.42	89.42	89.40	84.28[a]	**89.43**	87.44[a]	76.10[a]	86.62[a]	89.41
23	65.48[b]	55.47[a]	63.07[b]	60.30[b]	56.20[a]	63.71[b]	57.75	**66.92**[b]	56.05[a]	56.01[a]	56.15[a]	58.31
24	85.69[b]	*	78.37	77.81[a]	99.65[b]	81.18[b]	75.86[a]	**99.69**[b]	75.10[a]	99.65[b]	78.19	79.51
25	83.35[b]	63.81[a]	**94.23**[b]	85.53[b]	85.53[b]	85.53[b]	87.75[b]	75.09[a]	88.29[b]	78.60[b]	83.90[b]	75.61
Avg.	82.66	80.17	82.08	82.70	82.50	83.13	80.69	81.99	80.51	78.49	80.61	**83.29**
W/T/L	**17**/4/4	10/**12**/0	7/**15**/3	8/**14**/3	9/**14**/2	5/**15**/5	**16**/5/4	9/**13**/3	**16**/7/2	**18**/4/3	**14**/9/2	

Table 11 Comparison of balanced average accuracy of algorithms using C4.5 classifier

No.	CFS	ConsFS	mRMR	FCBF	ReliefF	MIFS	DISR	IWFS	mIMR	IAMB	JMI	SAFE
1	63.77^{a}	63.99^{a}	61.39^{a}	62.25^{a}	55.12^{a}	55.51^{a}	**65.38**^{b}	65.09	65.17	63.54^{a}	65.14	65.11
2	94.52	95.85^{a}	95.02^{a}	**96.19**^{b}	95.16	95.03	94.02^{a}	94.13^{a}	94.45^{a}	95.37	94.91	95.00
3	**96.61**	50.00^{a}	96.23^{a}	96.57	96.23^{a}	96.46	96.25^{a}	**96.61**	96.19^{a}	96.26^{a}	96.21^{a}	**96.61**
4	58.81	59.34^{b}	**59.39**^{b}	57.78^{a}	58.82	58.41^{a}	57.59^{a}	58.17^{a}	57.02^{a}	56.15^{a}	57.06^{a}	58.92
5	72.44	70.87^{a}	74.62^{b}	68.77^{a}	**77.10**^{b}	74.14^{b}	72.75	70.73^{a}	62.66^{a}	65.86^{a}	69.58^{a}	72.06
6	93.83^{a}	93.13^{a}	94.96^{b}	94.63	94.97^{b}	**95.12**^{b}	94.92	90.08^{a}	93.40^{a}	78.37^{a}	94.73	94.75
7	80.27	78.62^{a}	80.71	78.88^{a}	**82.41**^{b}	79.50^{a}	81.64	80.46	79.63^{a}	80.64	80.52	81.10
8	60.06^{a}	62.87	63.57^{b}	62.68	63.84^{b}	50.58^{a}	62.51	63.01	**64.44**^{b}	62.44	63.51^{b}	62.64
9	98.44^{a}	**99.99**^{b}	98.96	98.43^{a}	98.96	**99.99**^{b}	**99.99**^{b}	99.37^{b}	98.67^{a}	99.37^{b}	99.78^{b}	99.06
10	79.52	81.21^{b}	74.56^{a}	84.04^{b}	79.39^{a}	**84.47**^{b}	80.37^{b}	78.26^{a}	80.05^{b}	78.62^{a}	70.99^{a}	79.71
11	72.55^{a}	**73.86**^{b}	53.17^{a}	72.47^{a}	72.11^{a}	72.67^{a}	72.18^{a}	67.86^{a}	73.01^{a}	66.27^{a}	72.08^{a}	73.30
12	94.76^{b}	90.98^{a}	92.45^{a}	**95.20**^{b}	92.86^{a}	95.12^{b}	87.21^{a}	84.75^{a}	85.32^{a}	71.85^{a}	90.45^{a}	94.33
13	89.69^{a}	83.81^{a}	65.19^{a}	88.10^{a}	65.19^{a}	90.08	90.16	88.90^{a}	**90.64**^{b}	86.27^{a}	84.44^{a}	90.04
14	97.94^{a}	96.25^{a}	97.17^{a}	**99.99**^{b}	97.83^{a}	97.09^{a}	97.82^{a}	97.93^{a}	**99.99**^{b}	92.05^{a}	**99.99**^{b}	98.64
15	89.18^{a}	94.43^{b}	55.81^{a}	67.43^{a}	89.95^{a}	82.21^{a}	94.95^{b}	**97.74**^{b}	96.74^{b}	96.38^{b}	96.73^{b}	90.18
16	88.70^{a}	92.32^{b}	81.15^{a}	90.73^{a}	**92.37**^{b}	92.21^{b}	92.18	89.61^{a}	81.72^{a}	74.50^{a}	91.94	91.96
17	58.14^{a}	61.14	**62.10**	59.84^{a}	61.04	58.75^{a}	59.38^{a}	61.29	61.71	55.91^{a}	59.24^{a}	61.78
18	77.75	77.46	75.78^{a}	77.53	77.25	77.28	76.59^{a}	**78.06**	75.87^{a}	76.65^{a}	77.12	77.65
19	94.27^{a}	95.32	**95.41**	94.63^{a}	95.39	94.84^{a}	95.31	93.25^{a}	95.28	94.10^{a}	95.30	95.32
20	73.79^{a}	76.30^{a}	71.28^{a}	73.95^{a}	73.84^{a}	74.08^{a}	74.07^{a}	72.53^{a}	74.24^{a}	70.42^{a}	76.86^{a}	**77.41**

Table 11 continued

No.	CFS	ConsFS	mRMR	FCBF	ReliefF	MIFS	DISR	IWFS	mIMR	IAMB	JMI	SAFE
21	50.00	*	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00
22	82.30	*	50.00^{a}	72.31^{a}	72.98^{a}	82.74	82.87^{b}	82.16	89.48^{b}	73.73^{a}	$\mathbf{85.84}^{b}$	82.04
23	65.07	59.75^{a}	$\mathbf{68.50}^{b}$	68.20^{b}	58.66^{a}	67.05^{b}	63.27	60.83^{a}	59.13^{a}	57.20^{a}	61.32^{a}	64.00
24	$\mathbf{86.22}^{b}$	*	80.00	79.06^{a}	80.00	82.74	76.11^{a}	81.14	76.70^{a}	68.40^{a}	78.02^{a}	80.83
25	82.78^{b}	64.20^{a}	86.47^{b}	87.50^{b}	75.60	71.58^{a}	87.26^{b}	77.27^{b}	87.06^{b}	79.12^{b}	$\mathbf{88.58}^{b}$	75.64
Avg.	80.06	78.26	75.36	79.09	78.28	79.11	80.19	79.17	79.36	75.58	80.01	**80.32**
W/T/L	**12**/10/3	**12**/4/6	**13**/5/7	**14**/5/6	**11**/9/5	**11**/6/8	**10**/9/6	**13**/9/3	**14**/4/7	**18**/4/3	**11**/8/6	

a(*b*) represents the proposed method SAFE is significantly better (worse) than the other method at 5% significance level. Figures in bold indicates the result for the best classifier for that dataset. W/T/L denotes the number of datasets in which the proposed method SAFE wins, ties or loses with respect to the other feature selection method at 5% significance level after Benjamini–Hochberg adjustment for multiple comparisons. We use a computer with Windows 7, i5 processor, and the time-out threshold was kept at 30 mins. *Denotes solution not converged within threshold time

Table 12 Summary of wins/ties/loses (W/T/L) results for SAFE

Classifier	CFS	ConsFS	mRMR	FCBF	ReliefF	MIFS	DISR	IWFS	mIMR	IAMB	JMI
NB	**12**/9/4	**13**/3/6	**14**/5/6	**19**/2/4	**13**/5/7	**17**/5/3	**14**/6/5	**13**/7/3	**16**/1/8	**20**/3/2	**13**/6/6
LR	**12**/8/5	**13**/3/6	**13**/8/4	**14**/8/3	**12**/7/6	**11**/8/6	**13**/3/9	**14**/5/6	**13**/5/7	**18**/3/4	**10**/7/8
RDA	**11**/8/6	9/**11**/2	**13**/8/4	**10**/**10**/5	**11**/8/6	**13**/8/4	**11**/9/5	9/**12**/4	**13**/6/6	**13**/5/7	9/**11**/5
SVM	**17**/4/4	**15**/4/3	**17**/5/3	**17**/7/1	**13**/8/4	**14**/5/6	**12**/5/8	**19**/4/2	**14**/4/7	**15**/5/5	**10**/8/7
KNN	**17**/4/4	10/**12**/0	7/**15**/3	8/**14**/3	9/**14**/2	5/**15**/5	**16**/5/4	9/**13**/3	**16**/7/2	**18**/4/3	**14**/9/2
C4.5	**12**/10/3	**12**/4/6	**13**/5/7	**14**/5/6	**11**/9/5	**11**/6/8	**10**/9/6	**13**/9/3	**14**/4/7	**18**/4/3	**11**/8/6
Average	**14**/7/4	**12**/6/4	**13**/8/4	**14**/7/4	**12**/8/5	**12**/8/5	**13**/6/6	**13**/8/4	**14**/5/6	**17**/4/4	**11**/8/6

ConsFS method did not converge for 3 data sets within threshold time of 30 min, those results are not reported. Classifier-wise average is rounded to nearest whole number
Figures in bold denote the maximum of W/T/L

conditions. The difference, (W–L), is particularly large for 4 out of 6 classifiers in which cases SAFE wins in more than half of all data sets on average (NB = 60.3%, LR = 52.6%, SVM = 60%, C4.5 = 51.1%). SAFE is competitive with most feature selection methods in case of kNN, winning in 47.4% of all datasets on average. Compared to other methods, SAFE achieves the highest average accuracy in 3 out of 6 classifiers (NB, kNN, and C4.5), having maximum for kNN and minimum for RDA.

One general observation is that the performance of SAFE is better than that of redundancy-based methods on an average (in terms of W/T/L), which shows that complementary-based feature selection yields a more predictive subset of features. The fact that SAFE performs well across various classifiers and various domains, is a test of robustness of the proposed heuristic. For example, NB is a probabilistic Bayesian classifier that assumes conditional independence amongst the features given the class variable, whereas LR makes no assumptions about conditional independence of the features. SAFE performs well in either case, which demonstrates that the performance of SAFE does not degrade when features are not conditionally independent given the class, which is a limitation of CFS. CFS defeats SAFE in only 4 data sets when NB classifier is used. In fact, SAFE focuses on the interaction gain or loss, i.e., $I(F_i; F_j; Y) = I(F_i; F_j) - I(F_i; F_j|Y)$, that is the difference between unconditional and conditional dependence. However, in cases where all features are pairwise redundant, it does not do better than CFS.

Compared to the interaction-based feature selection methods such as DISR, IWFS, JMI, mIMR, SAFE does better in most cases, the reason being SAFE explicitly measures the redundancy and complementarity, unlike DISR, which takes the aggregate effect. JMI is the most competitive in terms of W/T/L results, though SAFE outperforms JMI in 11 out of 25 datasets on average. Our heurisic SAFE is a complementary-based feature selection criterion, and is expected to show superior performance when the data sets contain complementary features. To demonstrate this, we focus on two data sets—lung cancer and promoter—that are highly complementary as shown in Fig. 6. The results show that SAFE mostly wins or ties, and loses in very few cases for these two data sets.

To evaluate how well our proposed heuristic corresponds to the actual predictive accuracy, we examine a plot of predictive accuracy versus heuristic score for several experimental data sets. We split each data set into 70% training set and 30% test set as before, and randomly select 100 subsets of features of varying sizes ranging from 1 and n, where n is the total number of features in the data set. For each data set, the heuristic score is estimated using the training set, and the accuracy is determined using the test set. As an illustration, we present the results for two datasets 'Dermatology' and 'Cardiology' in Figs. 13 and 14 respectively, the former being predominantly redundant ($\alpha = 0.80$), and the latter being predominantly complementary ($\alpha = 0.21$). The plots show that, in general, there exists a correspondence between accuracy and score. As the score increases, the predictive accuracy also tends to increase in most cases.

7.4.2 Runtime comparison

Computation time is an important criterion that measures the time complexity of an algorithm. Table 13 shows the average execution time (in s) taken by each feature selection method. The results show that FCBF is the fastest of all, while ConsFS is the most expensive. ConsFS does not converge to solutions for three data sets with threshold time of 30 min. SAFE comes third after FCBF and IAMB, showing that our proposed heuristic is not very expensive in terms of time complexity. Though the worst case time complexity of SAFE is $O(n^2)$, in reality, the computation time for SAFE is acceptable.

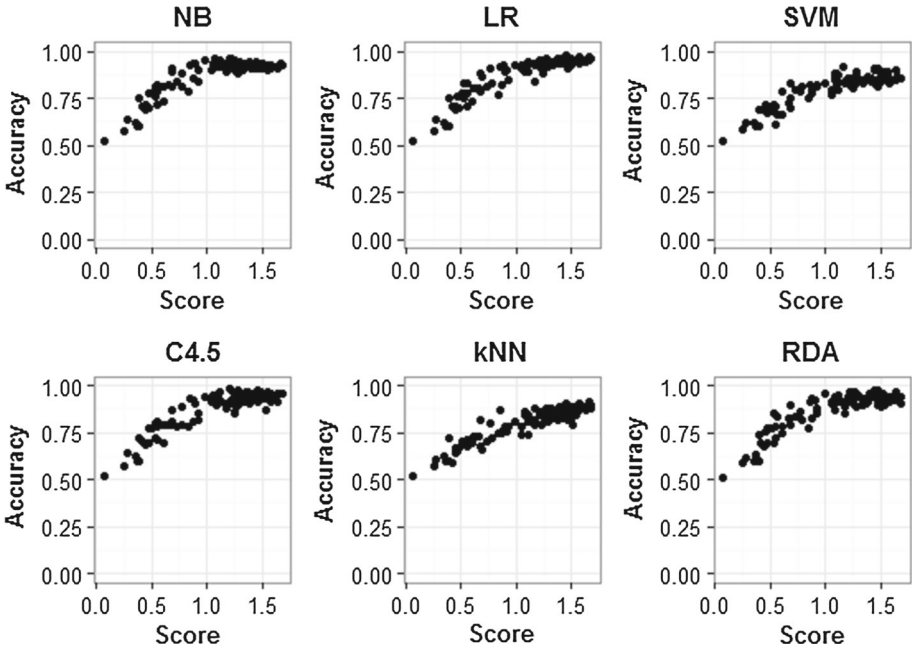

Fig. 13 Plot showing variation of predictive accuracy with heuristic score for dataset 'Dermatology' ($\alpha = 0.80$)

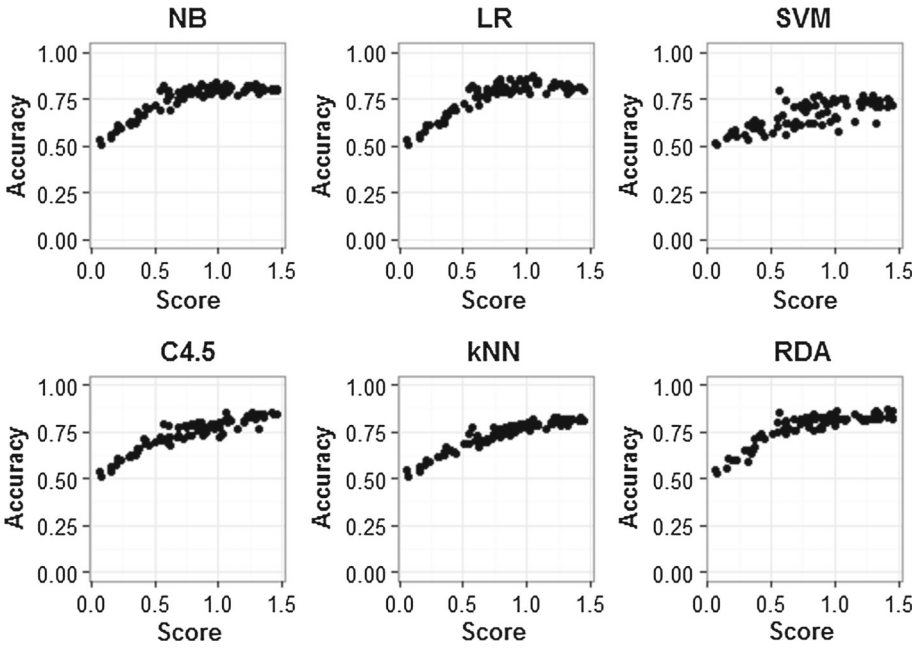

Fig. 14 Plot showing variation of predictive accuracy with heuristic score for dataset 'Cardiology' ($\alpha = 0.21$)

Table 13 Average execution time (s) for each iteration

Dataset	CFS	ConsFS	mRMR	FCBF	ReliefF	MIFS	DISR	IWFS	mIMR	IAMB	JMI	SAFE
Average	55.5	102.6	51.0	28.4	51.3	50.7	52.3	63.0	48.5	35.8	51.4	45.4

Table 14 Average number of features selected

Dataset	CFS	ConsFS	mRMR	FCBF	ReliefF	MIFS	DISR	IWFS	mIMR	IAMB	JMI	SAFE
Average	8.6	7.5	10.1	6.5	23.4	15.8	12.4	11.02	12.8	4.0	12.10	6.7

7.4.3 Number of selected features

Table 14 shows the average number of features selected by each feature selection method. IAMB selects the least number of features, while ReliefF selects the maximum number of features. SAFE selects an average of 6.7 features, which is the third lowest after IAMB and FCBF. MIFS, DISR, mIMR, JMI, and ReliefF select more features compared to SAFE. All methods have been able to remove a large number of irrelevant and redundant features.

8 Summary and conclusion

The main goal of feature selection is to find a small subset of features from the original ones such that it is highly predictive of the class. In this paper, we propose a filter-based feature selection criterion that relies on feature complementarity for searching an optimal subset. Unlike the existing redundancy-based methods, which only depend on relevance and redundancy, our proposed approach also aims to maximize complementarity. Incorporating feature complementarity as an additional search criterion enables us to leverage the power of complementarity, resulting in a smaller and more predictive subset of features. Since redundancy is generally modeled using feature correlation, the existing redundancy-based methods penalize all dependencies regardless of whether such dependence increases the predictive power or reduces it. As a result, our proposed approach is able to distinguish complementary feature subset from a subset of independent features, which the existing redundancy-based approach fails to do.

Using information theory framework, we explicitly measure complementarity, and integrate this in an adaptive evaluation criterion based on an interactive approach of multi-objective optimization. A feature of the proposed heuristic is that it adaptively optimizes relevance, redundancy, and complementarity while minimizing the subset size. Such an adaptive scoring criterion is new in feature selection. The proposed method not only helps to remove irrelevant and redundant features, but also selects complementary features, thus enhancing the predictive power of a subset. Using benchmark data sets and different classifiers, the experimental results show that the proposed method outperforms many existing methods for most data sets. The proposed method has acceptable time complexity, and effectively removes a large number of features. This paper shows that the proposed complementary-based feature selection method can be used to improve the classification performance of many popular classifiers for real life problems.

Acknowledgements This work is an outcome of the doctoral dissertation (Singha 2018). We are grateful to Suzanna Emelio of University of Kansas for proofreading this manuscript.

References

Abramson, N. (1963). *Information theory and coding*., McGraw-Hill electronic sciences series New York: McGraw-Hill.

Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning*, *6*(1), 37–66.

Almuallim, H., & Dietterich, T. G. (1991). Learning with many irrelevant features. In *AAAI* (Vol. 91, pp. 547–552).

Batista, G. E., Monard, M. C., et al. (2002). A study of *k*-nearest neighbour as an imputation method. *HIS*, *87*(251–260), 48.

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

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)*, *57*(1), 289–300.

Bontempi, G. & Meyer, P. E. (2010). Causal filter selection in microarray data. In *Proceedings of the 27th international conference on machine learning (ICML-10)* (pp. 95–102).

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. Boca Raton: CRC Press.

Brown, G. (2009). A new perspective for information theoretic feature selection. In D. van Dyk & M. Welling (Eds.), *Proceedings of the twelth international conference on artificial intelligence and statistics, proceedings of machine learning research* (Vol. 5, pp. 49–56).

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*(1), 27–66.

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

Chernbumroong, S., Cang, S., & Yu, H. (2015). Maximum relevancy maximum complementary feature selection for multi-sensor activity recognition. *Expert Systems with Applications*, *42*(1), 573–583.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory*. Hoboken, NJ: Wiley.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society, Series B*, *20*(2), 215–242.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge, NY: Cambridge University Press.

Dash, M., & Liu, H. (2003). Consistency-based search in feature selection. *Artificial Intelligence*, *151*(1–2), 155–176.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *39*(1), 1–38.

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

Doquire, G., & Verleysen, M. (2013). Mutual information-based feature selection for multilabel classification. *Neurocomputing*, *122*, 148–155.

Dougherty, J., Kohavi, R. & Sahami, M., et al. (1995). Supervised and unsupervised discretization of continuous features. In *Machine learning: proceedings of the 12th international conference* (Vol. 12, pp. 194–202).

Duch, W. (2006). Filter methods. In I. Guyon, M. Nikravesh, S. Gunn, & L. A. Zadeh (Eds.), *Feature extraction: Foundations and applications* (pp. 89–117). Berlin: Springer.

Fayyad, U. & Irani, K. B. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. In R. Bajcsy (Ed.), *Proceedings of the 13th international joint conference on artificial intelligence* (Vol. 2, pp. 1022–1029).

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

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association*, *84*(405), 165–175.

Gao, S., Ver Steeg, G., & Galstyan, A. (2016). Variational information maximization for feature selection. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 29, pp. 487–495). Red Hook: Curran Associates, Inc.,

Gheyas, I. A., & Smith, L. S. (2010). Feature subset selection in large dimensionality domains. *Pattern Recognition*, *43*(1), 5–13.
Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182.
Hall, M. A. (2000). Correlation-based feature selection for discrete and numeric class machine learning. In P. Langley (Ed.), *Proceedings of the 17th international conference on machine learning (ICML '00), Morgan Kaufmann, CA, USA* (pp. 359–366).
Hall, M. A., & Holmes, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data engineering*, *15*(6), 1437–1447.
Hastie, T., Tibshirani, R., Sherlock, G., Eisen, M., Brown, P. & Botstein, D. (1999). Imputing missing data for gene expression arrays.
Jakulin, A., & Bratko, I. (2003). Analyzing attribute dependencies. In N. Lavrač, D. Gamberger, L. Todorovski, & H. Blockeel (Eds.), *European conference on principles of data mining and knowledge discovery* (pp. 229–240). Berlin, Germany: Springer.
Jakulin, A. & Bratko, I. (2004). Testing the significance of attribute interactions. In C. E. Brodley (Ed.), *Proceedings of the 21st international conference on machine learning (ICML '04), ACM, NY, USA* (Vol. 69, pp. 52–59).
John, G. H., & Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. In P. Besnard & S. Hanks (Eds.), *Eleventh conference on uncertainty in artificial intelligence* (pp. 338–345). San Francisco, USA: Morgan Kaufmann.
Kira, K. & Rendell, L. A. (1992). A practical approach to feature selection. In *Proceedings of the 9th international workshop on machine learning* (pp. 249–256).
Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1), 273–324.
Kojadinovic, I. (2005). Relevance measures for subset variable selection in regression problems based on k-additive mutual information. *Computational Statistics & Data Analysis*, *49*(4), 1205–1227.
Koller, D. & Sahami, M. (1996). Toward optimal feature selection. In: L. Saitta (Ed.), *Proceedings of the 13th international conference on machine learning (ICML '96), Morgan Kaufmann, San Francisco, USA* (pp. 284–292).
Kononenko, I. (1994). Estimating attributes: Analysis and extensions of relief. In F. Bergadano & L. De Raedt (Eds.), *European conference on machine learning* (pp. 171–182). Springer.
Koprinska, I. (2009). Feature selection for brain-computer interfaces. In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 106–117). Springer.
Kotsiantis, S., & Kanellopoulos, D. (2006). Discretization techniques: A recent survey. *GESTS International Transactions on Computer Science and Engineering*, *32*(1), 47–58.
Krippendorff, K. (2009). Information of interactions in complex systems. *International Journal of General Systems*, *38*(6), 669–680.
Kwak, N., & Choi, C. H. (2002). Input feature selection for classification problems. *IEEE Transactions on Neural Networks*, *13*(1), 143–159.
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* (pp. 212–217).
Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data*. Hoboken: Wiley.
Liu, H. & Setiono, R. (1996). A probabilistic approach to feature selection—A filter solution. In L. Saitta (Ed.), *Proceedings of the 13th international conference on machine learning (ICML '96), Morgan Kaufmann, San Francisco, USA* (pp. 319–327).
Matsuda, H. (2000). Physical nature of higher-order mutual information: Intrinsic correlations and frustration. *Physical Review E*, *62*(3), 3096.
McGill, W. J. (1954). Multivariate information transmission. *Psychometrika*, *19*(2), 97–116.
Meyer, P. E., & Bontempi, G. (2006). On the use of variable complementarity for feature selection in cancer classification. In *Workshops on applications of evolutionary computation* (pp. 91–102). Berlin, Germany: Springer.
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.
Narendra, P. M., & Fukunaga, K. (1977). A branch and bound algorithm for feature subset selection. *IEEE Transactions on Computers*, *26*(9), 917–922.
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.
Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106.
R Core Team. (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/.

Rich, E., & Knight, K. (1991). *Artificial intelligence*. New York: McGraw-Hill.
Rissanen, J. (1986). Stochastic complexity and modeling. *The Annals of Statistics*, *14*(3), 1080–1100.
Robnik-Šikonja, M., & Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning*, *53*(1–2), 23–69.
Rosenthal, R. E. (1985). Concepts, theory, and techniques principles of multiobjective optimization. *Decision Sciences*, *16*(2), 133–152.
Roy, B. (1971). Problems and methods with multiple objective functions. *Mathematical Programming*, *1*(1), 239–266.
Ruiz, R., Riquelme, J. C., & Aguilar-Ruiz, J. S. (2002). Projection-based measure for efficient feature selection. *Journal of Intelligent & Fuzzy Systems*, *12*(3, 4), 175–183.
Saska, J. (1968). Linear multiprogramming. *Ekonomicko-Matematicky Obzor*, *4*(3), 359–373.
Senawi, A., Wei, H. L., & Billings, S. A. (2017). A new maximum relevance-minimum multicollinearity (MRmMC) method for feature selection and ranking. *Pattern Recognition*, *67*, 47–61.
Singha, S. (2018). Three essays on feature and model selection for classification and regression problems. Ph.D. thesis, School of Business, University of Kansas, Lawrence, KS.
Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., et al. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17*(6), 520–525.
Tsamardinos, I., Aliferis, C. F. & Statnikov, A. (2003). Algorithms for large scale Markov blanket discovery. In *Proceedings of the 16th international florida artificial intelligence research society conference (FLAIRS-03)* pp. 376–381.
Vergara, J. R., & Estévez, P. A. (2014). A review of feature selection methods based on mutual information. *Neural Computing and Applications*, *24*(1), 175–186.
Wang, G. & Lochovsky, F. H. (2004). Feature selection with conditional mutual information maximin in text categorization. In *Proceedings of the 13th ACM international conference on information and knowledge management, ACM* (pp. 342–349).
Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining: Practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann.
Yang, 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).
Yang, H. H. & Moody, J. (2000). Data visualization and feature selection: New algorithms for nongaussian data. In *Advances in neural information processing systems* (pp. 687–693).
Yaramakala, S. & Margaritis, D. (2005). Speculative Markov blanket discovery for optimal feature selection. In *Proceedings of the 5th IEEE international conference on data mining (ICDM-05), IEEE Computer Society, Washington, DC* (pp. 809–812).
Yeung, R. W. (1991). A new outlook on Shannon's information measures. *IEEE Transactions on Information Theory*, *37*(3), 466–474.
Yu, L. & Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In: T. Fawcett & N. Mishra (Eds.), *Proceedings of the 20th international conference of machine learning (ICML'03), AAAI Press, California, USA* (pp. 856–863).
Yu, L., & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, *5*(Oct), 1205–1224.
Zeng, Z., Zhang, H., Zhang, R., & Yin, C. (2015). A novel feature selection method considering feature interaction. *Pattern Recognition*, *48*(8), 2656–2666.
Zhang, Y., & Zhang, Z. (2012). Feature subset selection with cumulate conditional mutual information minimization. *Expert Systems with Applications*, *39*(5), 6078–6088.
Zhao, Z. & Liu, H. (2007). Searching for interacting features. In *Proceedings of the 20th international joint conference on artificial intelligence, Morgan Kaufmann Publishers Inc., San Francisco, USA* (pp. 1156–1161).