Neural Comput & Applic (2014) 24:175–186
DOI 10.1007/s00521-013-1368-0

INVITED REVIEW

A review of feature selection methods based on mutual information

Jorge R. Vergara · Pablo A. Estévez

Received: 15 February 2013 / Accepted: 21 February 2013 / Published online: 13 March 2013

Abstract In this work, we present a review of the state of the art of information-theoretic feature selection methods. The concepts of feature relevance, redundance, and complementarity (synergy) are clearly defined, as well as Markov blanket. The problem of optimal feature selection is defined. A unifying theoretical framework is described, which can retrofit successful heuristic criteria, indicating the approximations made by each method. A number of open problems in the field are presented.

Keywords Feature selection · Mutual information · Relevance · Redundancy · Complementarity · Sinergy · Markov blanket

1 Introduction

Feature selection has been widely investigated and used by the machine learning and data mining community. In this context, a feature, also called attribute or variable, represents a property of a process or system than has been measured or constructed from the original input variables. The goal of feature selection is to select the smallest feature subset given a certain generalization error, or alternatively finding the best feature subset with *k* features, that yields the minimum generalization error. Additional objectives of feature selection are as follows: (i) improve the generalization performance with respect to the model built using the whole set of features, (ii) provide a more robust generalization and a faster response with unseen data, and (iii) achieve a better and simpler understanding of the process that generates the data [23, 31]. We will assume that the feature selection method is used either as a preprocessing step or in conjunction with a learning machine for classification or regression purposes. Feature selection methods are usually classified in three main groups: wrapper, embedded, and filter methods [23]. Wrappers [31] use the induction learning algorithm as part of the function evaluating feature subsets. The performance is usually measured in terms of the classification rate obtained on a testing set, i.e., the classifier is used as a black box for assessing feature subsets. Although these techniques may achieve a good generalization, the computational cost of training the classifier a combinatorial number of times becomes prohibitive for high-dimensional datasets. In addition, many classifiers are prone to overlearning and show sensitiveness to initialization. Embedded methods [38] incorporate knowledge about the specific structure of the class of functions used by a certain learning machine, e.g., bounds on the leave-one-out error of SVMs [64]. Although usually less computationally expensive than wrappers, embedded methods still are much slower than filter approaches, and the features selected are dependent on the learning machine. Filter methods [17] assume complete independence between the learning machine and the data, and therefore use a metric independent of the induction learning algorithm to assess feature subsets. Filter methods are relatively robust against overfitting, but may fail to select the best feature subset for classification or regression. In the literature, several criteria

J. R. Vergara
Department of Electrical Engineering,
Faculty of Physical and Mathematical Sciences,
University of Chile, Santiago, Chile
e-mail: jorgever@ing.uchile.cl

P. A. Estévez (✉)
Department of Electrical Engineering and Advanced Mining
Technology Center, Faculty of Physical and Mathematical
Sciences, University of Chile, Santiago, Chile
e-mail: pestevez@ing.uchile.cl

have been proposed to evaluate single features or feature subsets, among them: inconsistency rate [28], inference correlation [44], classification error [18], fractal dimension [45], distance measure [8, 50], etc. Mutual information (MI) is a measure of statistical independence that has two main properties. First, it can measure any kind of relationship between random variables, including nonlinear relationships [14]. Second, MI is invariant under transformations in the feature space that are invertible and differentiable, e.g., translations, rotations, and any transformation preserving the order of the original elements of the feature vectors [35, 36]. Many advances in the field have been reported in the last 20 years since the pioneer work of Battiti [4]. Battiti defined the problem of feature selection as the process of selecting the k most relevant variables from an original feature set of m variables, $k < m$. Battiti proposed the greedy selection of a single feature at a time, as an alternative to evaluate the combinatorial explosion of all feature subsets belonging to the original set. The main assumptions of Battiti's work were the following: (a) features are classified as relevant and redundant; (b) an heuristic functional is used to select features, which allows controlling the tradeoff between relevancy and redundancy; (c) a greedy search strategy is used; and (d) the selected feature subset is assumed optimal. These four assumptions will be revisited in this work to include recent work on (a) new definitions on relevant features and other types of features, (b) new information-theoretic functional derived from first principles, (c) new search strategies, and (d) new definitions of optimal feature subset. In this work, we present a review of filtering feature selection methods based on mutual information, under a unified theoretical framework. We show the evolution of feature selection methods on the last 20 years, describing advantages and drawbacks. The remainder of this work is organized as follows. In Sect. 2, a background on MI is presented. In Sect. 3, the concepts of relevant, redundant, and complementary features are defined. In Sect. 4, the problem of optimal feature selection is defined. In Sect. 5, a unified theoretical framework is presented, which allows us to show the evolution of different MI feature selection methods, as well as their advantages and drawbacks. In Sect. 6, a number of open problems in the field are presented. Finally, in Sect. 7, we present the conclusions of this work.

2 Background on MI

2.1 Notation

In this work, we will use only discrete random variables, because in practice the variables used in most feature selection problems are either discrete by nature or by quantization. Let F be a feature set and C an output vector representing the classes of a real process. Let us assume that F is the realization of a random sampling of an unknown distribution, where f_i is the ith variable of F and $f_i(j)$ is the jth sample of vector f_i. Likewise, c_i is the ith component of C and $c_i(j)$ is the jth sample of vector c_i. Uppercase letters denote random sets of variables, and lowercase letters denote individual variables from these sets.

Other notations and terminologies used in this work are the following:

S	Subset of current selected variables.
f_i	Candidate feature to be added to or deleted from the subset of selected features S.
$\{f_i, f_j\}$	Subset composed of the variables f_i and f_j.
$\neg f_i$	All variables in F except f_i. $\neg f_i = F \setminus f_i$.
$\{f_i, S\}$	Subset composed of variable f_i and subset S.
$\neg\{f_i, S\}$	All variables in F except the subset $\{f_i, S\}$. $\neg\{f_i, S\} = F \setminus \{f_i, S\}$
$p(f_i, C)$	Joint mass probability between variables f_i and C.
$\lvert \cdot \rvert$	Absolute value/cardinality of a set.

The sets mentioned above are related as follows: $F = f_i \cup S \cup \neg\{f_i, S\}, \emptyset = f_i \cap S \cap \neg\{f_i, S\}$. The number of samples in F is n and the total number of variables in F is m.

2.2 Basic definitions

Entropy, divergence, and mutual information are basic concepts defined within information theory [14]. In its origin, information theory was used within the context of communication theory, to find answers about data compression and transmission rate [52]. Since then, information theory principles have been largely incorporated into machine learning, see for example Principe [47].

2.2.1 Entropy

Entropy (H) is a measure of uncertainty of a random variable. The uncertainty is related to the probability of occurrence of an event. Intuitively, high entropy means that each event has about the same probability of occurrence, while low entropy means that each event has a different probability of occurrence. Formally, the entropy of a discrete random variable x, with mass probability $p(x(i)) = Pr\{x = x(i)\}, x(i) \in x$ is defined as:

$$H(x) = -\sum_{i=1}^{n} p(x(i)) \log_2(p(x(i))). \tag{1}$$

Entropy is interpreted as the expected value of the negative of the logarithm of mass probability. Let x and y be two random discrete variables. The joint entropy of

x and y, with joint mass probability $p(x(i), y(j))$, is the sum of the uncertainty contained by the two variables. Formally, joint entropy is defined as follows:

$$H(\{x,y\}) = -\sum_{i=1}^{n}\sum_{j=1}^{n} p(x(i),y(j)) \cdot \log_2(p(x(i),y(j))). \tag{2}$$

The joint entropy has values in the range,

$$\max(H(x),H(y)) \leq H(\{x,y\}) \leq H(x) + H(y). \tag{3}$$

The maximum value in inequality (3) happens when x and y are completely independent. The minimum value occurs when x is completely dependent on y. The conditional entropy measures the remaining uncertainty of the random variable x when the value of the random variable y is known. The minimum value of the conditional entropy is zero, and it happens when x is statistically dependent on y, i.e., there is no uncertainty in x if we know y. The maximum value happens when x and y are statistically independent, i.e., the variable y does not add information to reduce the uncertainty of x. Formally, the conditional entropy is defined as:

$$H(x|y) = \sum_{j=1}^{n} p(y(j)) \cdot H(x|y = y(j)) \tag{4}$$

where

$$0 < H(x|y) < H(x), \tag{5}$$

and $H(x|y = y(j))$ is the entropy of all $x(i)$, which are associated with $y = y(j)$.

Another way of representing the conditional entropy is:

$$H(x|y) = H(\{x,y\}) - H(y). \tag{6}$$

2.2.2 Mutual information

The mutual information (MI) is a measure of the amount of information that one random variable has about another variable [14]. This definition is useful within the context of feature selection because it gives a way to quantify the relevance of a feature subset with respect to the output vector C. Formally, the MI is defined as follows:

$$I(x;y) = \sum_{i=1}^{n}\sum_{j=1}^{n} p(x(i),y(j)) \cdot \log\left(\frac{p(x(i),y(j))}{p(x(i)) \cdot p(y(j))}\right), \tag{7}$$

where MI is zero when x and y are statistically independent, i.e., $p(x(i),y(j)) = p(x(i)) \cdot p(y(j))$. The MI is related linearly to entropies of the variables through the following equations:

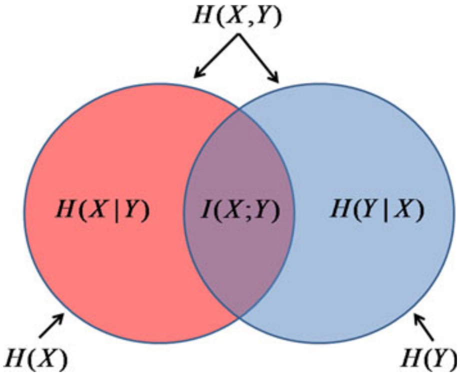

Fig. 1 Venn diagram showing the relationships between MI and entropies

$$I(x;y) = \begin{cases} H(x) - H(x|y) \\ H(y) - H(y|x) \\ H(x) + H(y) - H(x,y). \end{cases} \tag{8}$$

Figure 1 shows a Venn diagram with the relationships described in (8).

Let z be a discrete random variable. Its interaction with the other two variables $\{x, y\}$ can be measured by the conditional MI, which is defined as follows:

$$I(x;y|z) = \sum_{i=1}^{n} p(z(i)) I(x;y|z = z(i)), \tag{9}$$

where $I(x;y|z = z(i))$ is the MI between x and y in the context of $z = z(i)$. The conditional MI allows measuring the information of two variables in the context of a third one, but it does not measure the information among the three variables. Multi-information is an interesting extension of MI, proposed by McGill [42], which allows measuring the interaction among more than two variables. For the case of three variables, the multi-information is defined as follows:

$$I(x;y;z) = \begin{cases} I(\{x,y\};z) - I(x;z) - I(y;z) \\ I(y;z|x) - I(y;z). \end{cases} \tag{10}$$

The multi-information is symmetrical, i.e., $I(x;y;z) = I(x;z;y) = I(z;y;x) = I(y;x;z) = \ldots$ The multi-information has not been widely used in the literature, due to its difficult interpretation, e.g., the multi-information can take negative values, among other reasons. However, there are some interesting papers about the interaction among variables that use this concept [5, 30, 42, 68]. The multi-information can be understood as the amount of information common to all variables (or set of variables), but that is not present in any subset of these variables. To better understand the concept of multi-information within the context of feature selection, let us consider the following example.

Example 1 Let x_1, x_2, x_3 be independent binary random variables. The output of a given system is built through the function $C = x_1 + (x_2 \oplus x_3)$, and $x_4 = x_1$, where + stands for the OR logic function and $\oplus$ represents the XOR logic function.

x_1	x_2	x_3	x_4	$x_2 \oplus x_3$	C
0	0	0	0	0	0
1	0	0	1	0	1
0	1	0	0	1	1
1	1	0	1	1	1
0	0	1	0	1	1
1	0	1	1	1	1
0	1	1	0	0	0
1	1	1	1	0	1

Using Eq. (10) to measure the multi-information among x_2, x_3, and C gives: $I(x_2;x_3;C) = I(\{x_2,x_3\};C) - I(x_2;C) - I(x_3;C)$. Notice that the relevance of single features x_2 and x_3 with respect to C is null, since $I(x_2;C) = I(x_3;C) = 0$, but the joint information of $\{x_2,x_3\}$ with respect to C is greater than zero, $I(\{x_2,x_3\};C) > 0$. In this case, x_2 and x_3 interact positively to predict C, and this yields a positive value of the multi-information among these variables. The multi-information among the variables x_1, $x_{4,}$ and C is given by: $I(x_1;x_4;C) = I(\{x_1,x_4\};C) - I(x_1;C) - I(x_4;C)$. The relevance of individual features x_1 and x_4 is the same, i.e., $I(x_1;C) = I(x_4;C) > 0$. In this case, the joint information provided by x_1 and x_4 with respect to C is the same as that of each variable acting separately, i.e., $I(\{x_1,x_4\};C) = I(x_1;C) = I(x_4;C)$. This yields a negative value of the multi-information among these variables. We can deduce that the interaction between x_1 and x_4 does not provide any new information about C. Let us consider now the multi-information among x_1, x_2 and C, which is zero: $I(x_1;x_2;C) = I(\{x_1,x_2\};C) - I(x_1;C) - I(x_2;C) = 0$. Since feature x_2 only provides information about C when interacting with x_3, then $I(\{x_1,x_2\};C) = I(x_1;C)$. In this case, features x_1 and x_2 do not interact in the knowledge of C.

From the viewpoint of feature selection, the value of the multi-information (positive, negative, or zero) gives rich information about the kind of interaction there is among the variables. Let us consider the case where we have a set of already selected features S and a candidate feature f_i, and we measure the multi-information of these variables with the class variable C, $I(f_i;S;C) = I(S;C|f_i) - I(S;C)$. When the multi-information is positive, it means that feature f_i and S are complementary. On the other hand, when the multi-information is negative, it means that by adding f_i, we are diminishing the dependence between S and C, because f_i and S are redundant. Finally, when the multi-information is zero, it means that f_i is irrelevant with respect to the dependency between S and C.

The mutual information between a set of m features and the class variable C can be expressed compactly in terms of multi-information as follows:

$$I(\{x_1, x_2, \ldots, x_m\}; C) = \sum_{k=1}^{m} \sum_{\substack{\forall S \subseteq \{x_1,\ldots,x_m\} \\ |S|=k}} I([S \cup C]), \tag{11}$$

where $I([S \cup C]) = I(s_1; s_2; \cdots; s_k; C)$. Note that the sum on the right side of Eq. (11) is taken over all subsets S of size k drawn from the set $\{x_1, \ldots, x_m\}$.

3 Relevance, redundancy, and complementarity

The filter approach to feature selection is based on the idea of relevance, which we will explore in more detail in this section. Basically, the problem is to find the feature subset of minimum cardinality that preserves the information contained in the whole set of features with respect to C. This problem is usually solved by finding the relevant features and discarding redundant and irrelevant features. In this section, we review the different definitions of relevance, redundancy, and complementarity found in the literature.

3.1 Relevance

Intuitively, a given feature is relevant when either individually or together with other variables, it provides information about C. In the literature, there are many definitions of relevance, including different levels of relevance [1, 2, 4, 6, 10, 15, 23, 31, 46, 67] used a probabilistic framework to define three levels of relevance: strongly relevant, weakly relevant, and irrelevant features, as shown in Table 1. Strongly relevant features provide unique information about C, i.e., they cannot be replaced by other features. Weakly relevant features provide information about C, but they can be replaced by other features without losing information about C. Irrelevant features do not provide information about C, and they can be discarded without losing information. A drawback of the probabilistic approach is the need of testing the conditional independence for all possible feature subsets and estimating the probability density functions (pdfs) [48].

An alternative definition of relevance is given under the framework of mutual information [6, 21, 32, 33, 37, 53, 55, 67]. An advantage of this approach is that there are several good methods for estimating MI. The last column of

Table 1 Levels of relevance for candidate feature f_i, according to probabilistic framework [31] and mutual information framework [43]

Relevance level	Condition	Probabilistic approach	Mutual information approach
Strongly relevant	$\nexists$	$p(C \vert f_i, \neg f_i) \neq p(C \vert \neg f_i)$	$I(f_i; C \vert \neg f_i) > 0$
Weakly relevant	$\exists S \subset \neg f_i$	$p(C \vert f_i, \neg f_i) = p(C \vert \neg f_i)$ $\wedge p(C \vert f_i, S) \neq p(C \vert S)$	$I(f_i; C \vert \neg f_i) = 0$ $\wedge$ $I(f_i; C \vert S) > 0$
Irrelevant	$\forall S \subseteq \neg f_i$	$p(C \vert f_i, S) = p(C \vert S)$	$I(f_i; C \vert S) = 0$

Table 1 shows how the three levels of individual relevance are defined in terms of MI.

The definitions shown in Table 1 give rise to several drawbacks, which are summarized as follows:

1. To classify a given feature f_i, as irrelevant, it is necessary to assess all possible subsets S of $\neg f_i$. Therefore, this procedure is subject to the curse of dimensionality [7, 57].
2. The definition of strongly relevant features is too restrictive. If two features provide information about the class but are redundant, then both features will be discarded by this criterion. For example, let $\{x_1, x_2, x_3\}$ be a set of 3 variables, where $x_1 = x_2$, and x_3 is noise, and the output class is defined as $C = x_1$. Following the strong relevance criterion, we have $I(x_1;C|\{x_2, x_3\}) = I(x_2;C|\{x_1, x_3\}) = I(x_3;C|\{x_1, x_2\}) = 0$.
3. The definition of weak relevance is not enough for deciding whether to discard a feature from the optimal feature set. It is necessary to discriminate between redundant and non-redundant features.

3.2 Redundancy

Yu and Liu [67] proposed a finer classification of features into weakly relevant but redundant and weakly relevant but non-redundant. Moreover, the authors defined the set of optimal features as the one composed by strongly relevant features and weakly relevant but non-redundant features. The concept of redundancy is associated with the level of dependency among two or more features. In principle, we can measure the dependency of a given feature f_i with respect to a feature subset $S \subseteq \neg f_i$, by simply using the MI, $I(f_i;S)$. This information-theoretic measure of redundancy satisfies the following properties: it is symmetric, nonlinear, nonnegative, and does not diminish when adding new features [43]. However, using this measure, it is not possible to determine concretely with which features of S is f_i redundant. This calls for more elaborated criteria of redundancy, such as the Markov blanket [33, 67] and total correlation [62]. The Markov blanket is a strong condition for conditional independence and is defined as follows.

Definition 1 *(Markov blanket)* Given a feature f_i, the subset $M \subseteq \neg f_i$ is a Markov blanket of f_i iff [33, 67]:

$$p(\{F \backslash \{f_i, M\}, C\}|\ \{f_i, M\}) = p(\{F \backslash \{f_i, M\}, C\}|\ M). \tag{12}$$

This condition requires that M subsumes all the information that f_i has about C, but also about all other features $\{F \backslash \{f_i,M\}\}$. It can be proved that strongly relevant features do not have a Markov blanket [67].

The Markov blanket condition given by Eq. (12) can be rewritten in the context of information theory as follows [43]:

$$I(f_i; \{C, \neg f_i, M\}|\ M) = 0. \tag{13}$$

An alternative measure of redundancy is the total correlation or multivariate correlation [62]. Given a set of features $F = \{f_1, \ldots, f_m\}$, the total correlation is defined as follows:

$$C(f_1; \ldots; f_m) = \sum_{i=1}^{m} H(f_i) - H(f_1, \ldots, f_m). \tag{14}$$

Total correlation measures the common information (redundancy) among all the variables in F. If we want to measure the redundancy between a given variable f_i and any feature subset $S \subseteq \neg f_i$, then we can use the total correlation as:

$$C(f_i; S) = H(f_i) + H(S) - H(f_i, S); \tag{15}$$

however, this corresponds to the classic definition of MI, i.e., $C(f_i;S) = I(f_i;S)$.

3.3 Complementarity

The concept of complementarity has been re-discovered several times [9, 10, 12, 43, 61]. Recently, it has become more relevant because of the development of more efficient techniques to estimate MI in high-dimensional spaces [27, 34]. Complementarity, also known as synergy, measures the degree of interaction between an individual feature f_i and feature subset S given C, through the following expression $(I(f_i;S|C))$. To illustrate the concept of

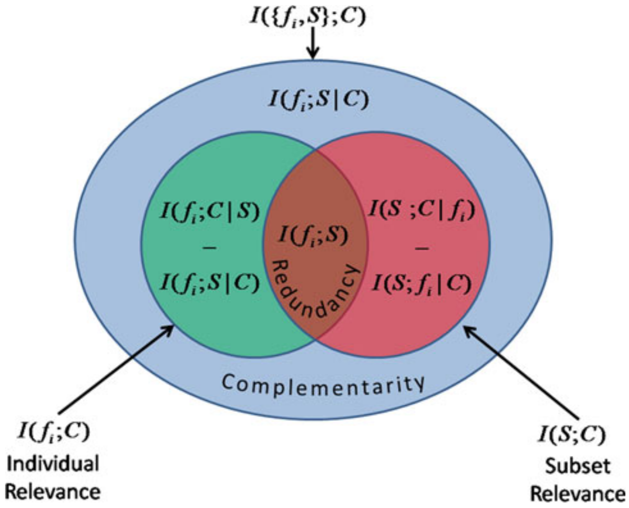

Fig. 2 Venn diagram showing the relationships among complementarity, redundancy, and relevancy, assuming that the multi-information among f_i, S and C is positive

complementarity, we will start expanding the multi-information among f_i, C and S. Decomposing the multi-information in its three possible expressions, we have:

$$I(f_i;S;C) = \begin{cases} I(f_i;S|C) - I(f_i;S) \\ I(f_i;C|S) - I(f_i;C) \\ I(S;C|f_i) - I(S;C). \end{cases} \quad (16)$$

According to Eq. (16), the first row shows that the multi-information can be expressed as the difference between complementarity ($I(f_i;S|C)$) and redundancy ($I(f_i;S)$). A positive value of the multi-information entails a dominance of complementarity over redundancy. Analyzing the second row of Eq. (16), we observe that this expression becomes positive when the information that f_i has about C is greater when it interacts with subset S with respect to the case when it does not. This effect is called complementarity. The third row of Eq. (16) gives us another viewpoint of the complementarity effect. The multi-information is positive when the information that S has about C is greater when it interacts with feature f_i compared to the case when it does not interact. Assuming that the complementarity effect is dominant over redundancy, Fig. 2 illustrates a Venn diagram with the relationships among complementarity, redundancy, and relevancy.

4 Optimal feature subset

In this section, we review the different definitions of the optimal feature subset, S_{opt}, given in the literature, as well as the search strategies used for obtaining this optimal set. According to [58], in practice, the feature selection problem must include a classifier or an ensemble of classifiers, and a performance metric. The optimal feature subset is defined as the one that maximizes the performance metric having minimum cardinality. However, filter methods are independent of both the learning machine and the performance metric. Any filter method corresponds to a definition of relevance that employs only the data distribution [58]. Yu and Liu [67] defined the optimal feature set as composed of all strongly relevant features and the weakly relevant but not redundant features. In this section, we review the definitions of the optimal feature subset from the viewpoint of filter methods, in particular MI feature selection methods. The key notion is conditional independence, which allows defining the sufficient feature subset as follows [6, 24]:

Definition 2 $S \subseteq F$ is a sufficient feature subset iff

$$p(C|F) = p(C|S). \quad (17)$$

This definition implies that C and $\neg S$ are conditionally independent, i.e., $\neg S$ provides no additional information about C in the context of S. However, we still need a search strategy to select the feature subset S, and an exhaustive search using this criterion is impractical due to the curse of dimensionality.

In probability, the measure of sufficient feature subset can be expressed as the expected value over $p(F)$ of the Kullback–Leibler divergence between $p(C|F)$ and $p(C|S)$ [33]. According to Guyon et al. [24], this can be expressed in terms of MI as follows:

$$\mathrm{DMI}(S) = I(F;C) - I(S;C). \quad (18)$$

Guyon et al. [24] proposed solving the following optimization problem:

$$\min_{S \subseteq F} |S| + \lambda \cdot \mathrm{DMI}(S), \quad (19)$$

where $\lambda > 0$ represents the Lagrange multiplier. If S is a sufficient feature subset, then $\mathrm{DMI}(S) = 0$ and Eq. (19) is reduced to $\min_{S \subseteq F} |S|$. Since $I(F;C)$ is constant, Eq. (19) is equivalent to:

$$\min_{S \subseteq F} |S| - \lambda \cdot I(S;C). \quad (20)$$

The feature selection problem corresponds to finding the smallest feature subset that maximizes $I(S;C)$. Since the term $\min_{S \subseteq F} |S|$ is discrete, the optimization of (20) is difficult. Tishby et al. [55] proposed replacing the term $\min_{S \subseteq F} |S|$ with $I(F;S)$.

An alternative approach to optimal feature subset selection is using the concept of the Markov blanket (MB). Remember that the Markov blanket, M, of a target variable C, is the smallest subset of F such that C is independent of the rest of the variables $F \backslash M$. Koller and Sahami [33] proposed using MBs as the basis for feature elimination.

They proved that features eliminated sequentially based on this criterion remain unnecessary. However, the time needed for inducing an MB grows exponentially with the size of this set when considering full dependencies. Therefore, most MB algorithms implement approximations based on heuristics, e.g., finding the set of k features that are strongly correlated with a given feature [33]. Fast MB discovery algorithms have been developed for the case of distributions that are faithful to a Bayesian Network [58, 59]. However, these algorithms require that the optimal feature subset does not contain multivariate associations among variables, which are individually irrelevant but become relevant in the context of others [11]. In practice, this means for example that current MB discovery algorithms cannot solve Example 1 due to the XOR function.

An important caveat is that both feature selection approaches, sufficient feature subset and MBs, are based on estimating the probability distribution of C given the data. Estimating posterior probabilities is a harder problem than classification, e.g., in using a 0\ 1-loss function only the most probable classification is needed. Therefore, this effect may render some features contained in sufficient feature subset or in the MB of C unnecessary [24, 56, 58].

4.1 Relationship between MI and Bayes error classification

There are some interesting results relating the MI between a random discrete variable f and a random discrete target variable C, with the minimum error obtained by maximum a posteriori classifier (Bayer classification error) [14, 20, 26]. The Bayes error is bounded above and below according to the following expression:

$$1-\frac{I(f;C)+\log(2)}{\log(|C|)} \leq e_{\text{bayes}}(f) \leqslant \frac{1}{2}(H(C)-I(f;C)). \tag{21}$$

Interestingly, Eq. (21) shows that both limits are minimized when the MI, $I(f;C)$ is maximized.

4.2 Search strategies

According to Guyon et al. [24], a feature selection method has three components: (1) evaluation criterion definition, e.g., relevance for filter methods, (2) evaluation criterion estimation, e.g., sufficient feature selection or MB for filter methods, and (3) search strategies for feature subset generation. In this section, we briefly review the main search strategies used by MI feature selection methods. Given a feature set F of cardinality m, there are 2^m possible subsets; therefore, an exhaustive search is impractical for high-dimensional datasets.

There are two basic search strategies: optimal methods and sub-optimal methods [63]. Optimal search strategies include exhaustive search and accelerated methods based on the monotonic property of a feature selection criterion, such as branch and bound. But optimal methods are impractical for high-dimensional datasets; therefore, sub-optimal strategies must be used.

Most popular search methods are sequential forward selection (SFS) [65] and sequential backward elimination (SBE) [41]. Sequential forward selection is a bottom–up search, which starts with an empty set and adds new features one at a time. Formally, it adds the candidate feature f_i that maximizes $I(S;C)$ to the subset of selected features S, i.e.,

$$S = S \cup \{\arg\max_{f_i \in F \setminus S}(I(\{S,f_i\};C))\}. \tag{22}$$

Sequential backward elimination is a top–down approach, which starts with the whole set of features and deletes one feature at a time. Formally, it starts with $S = F$ and proceeds deleting the less informative features one at a time, i.e,

$$S = S \setminus \{\arg\min_{f_i \in S}(I(\{S \setminus f_i\};C)\}. \tag{23}$$

Usually, backward elimination is computationally more expensive than forward selection, e.g., when searching for a small subset of features. However, backward elimination can usually find better feature subsets, because most forward selection methods do not take into account the relevance of variables in the context of features not yet included in the subset of selected features [23]. Both kinds of searching methods suffer from the nested effect, meaning that in forward selection, a variable cannot be deleted from the feature set once it has been added, and in backward selection, a variable cannot be reincorporated once it has been deleted. Instead of adding a single feature at a time, some generalized forward selection variants add several features, to take into account the statistical relationship between variables [63]. Likewise, the generalized backward elimination deletes several variables at a time. An enhancement may be obtained by combining forward and backward selection, avoiding the nested effect. The strategy “plus-l-take-away-r” [54] adds to S l features and then removes the worst r features if $l > r$, or deletes r features and then adds l features if $r < l$.

5 A unified framework for mutual information feature selection

Many MI feature selection methods have been proposed in the last 20 years. Most methods define heuristic functionals to assess feature subsets combining definitions of relevant and redundant features. Brown et al. [10] proposed a unifying framework for information-theoretic feature selection methods. The authors posed the feature selection problem as a conditional likelihood of the class labels, given features. Under the filter assumption [10], conditional likelihood is equivalent to conditional mutual information (CMI), i.e., the feature selection problem can be posed as follows:

$$\begin{aligned}&\min_{S\subseteq F}|S|\\&\text{subject to}:\ \min_{S\subseteq F} I(\neg S;C|S).\end{aligned}\tag{24}$$

This corresponds to the smallest feature subset such that the CMI is minimal. Starting from this objective function, the authors used MI properties to deduce some common heuristic criteria used for MI feature selection. Several criteria can be unified under the proposed framework. In particular, they showed that common heuristics based on linear combinations of information terms, such as Battiti's MIFS [4], conditional infomax feature extraction (CIFE) [22, 40], minimum-redundancy maximum relevance (mRMR) [46], and joint mutual information (JMI) [66], are all low-order approximations to the conditional likelihood optimization problem. However, the unifying framework proposed by Brown et al. [10] fell short of deriving (explaining) nonlinear criteria using min or max operators such as Conditional Mutual Information Maximization (CMIM) [21], Informative Fragments [61], and ICAP [29].

Let us start with the assumption that $I(F;C)$ measures all the information about the target variable contained in the set of features. This assumption is based on the additivity property of MI [14, 32], which states that the information about a given system is maximal when all features (F) are used to estimate the target variable (C). Using the chain rule, $I(F;C)$ can be decomposed as follows:

$$I(F;C) = I(S;C) + I(\neg S;C|S).\tag{25}$$

As $I(F;C)$ is constant, maximizing $I(S;C)$ is equivalent to minimizing $I(\neg S;C|S)$. Many MI feature selection methods maximize the first term on the right side of (25). This is known as the criterion of maximal dependency (MD) [46]. On the other hand, other criteria are based on the idea of minimizing the CMI, i.e., the second term on the right-hand side of Eq. (25).

In the following, we describe the approach of Brown et al. [10] for deriving sequential forward selection and sequential backward elimination algorithms, which are based on minimizing the CMI. For the convenience of the reader, we present the equivalent procedure in parallel when maximizing dependency (MD). In practice, a search strategy is needed to find the best feature subset. As we saw in Sect. 4.2, the most popular methods are sequential forward selection and sequential backward elimination. Before proceeding, we need to define some notation.

S^t Subset of selected variables at time t.

f_i Candidate feature to be added to or eliminated from feature subset S^t at time t.

$$f_i = \arg\max_{f_i\in\neg S^t} I(f_i;C|S^t) \quad \text{in forward selection.}$$
$$f_i = \arg\min_{f_i\in S^t} I(f_i;C|S^t\backslash f_i) \quad \text{in backward elimination.}$$

s_j A given feature in S^t.

$\neg s_j$ The complement set of feature s_j with set S^t, i.e., $\neg s_j = S^t\backslash s_j$

S^{t+1} Subset of selected variables at time t+1.

$$S^{t+1} \leftarrow \{S^t, f_i\} \quad \text{in forward selection.}$$
$$S^{t+1} \leftarrow S^t\backslash f_i \quad \text{in backward elimination.}$$

$\neg S^{t+1}$ Complement of feature subset S^{t+1}, i.e., $F = \{S^{t+1}, \neg S^{t+1}\}$.

$$\neg S^{t+1} \leftarrow \{\neg S^t\backslash f_i\} \quad \text{in forward selection.}$$
$$\neg S^{t+1} \leftarrow \{\neg S^t, f_i\} \quad \text{in backward elimination.}$$

Table 2 shows that for the case of sequential forward selection, we achieve the same result when using the MD or CMI approach: the SFS algorithm consists of maximizing $I(f_i;C|S^t)$. Analogously, Table 3 shows that for the case of sequential backward elimination, again we achieve the same result when using MD or CMI approaches: the SBE algorithm consists of minimizing $I(f_i;C|S^t\backslash f_i)$.

For space limitations, we will develop here only the case of forward feature selection, but the procedure is analogous for the case of backward feature elimination. The expression $I(f_i;C|S^t)$ can be expanded as follows [12]:

$$I(f_i;C|S^t) = I(f_i;C) - I(f_i;S^t) + I(f_i;S^t|C).\tag{26}$$

Table 2 Parallel between MD and CMI approaches for sequential forward selection

MD	CMI
$\max_{f_i\in\neg S^t} I(S^{t+1};C) =$	$\min_{f_i\in\neg S^t} I(\neg S^{t+1};C\|S^{t+1})$
$\max_{f_i\in\neg S^t} I(\{S^t,f_i\};C) =$	$\min_{f_i\in\neg S^t} I(\neg S^t\backslash f_i;C\|\{S^t,f_i\})$
$\max_{f_i\in\neg S^t} I(S^t;C)^a + \max_{f_i\in\neg S^t} I(f_i;C\|S^t)$	$\min_{f_i\in\neg S^t} I(\neg S^t;C\|S^t)^b + \min_{f_i\in\neg S^t}(-I(f_i;C\|S^t))$
$\Downarrow$	$\Downarrow$
$\max_{f_i\in\neg S^t} I(f_i;C\|S^t)$	$\max_{f_i\in\neg S^t} I(f_i;C\|S^t)$

[a] This term is independent of f_i

[b] This term has the same value $\forall\, f_i$

Table 3 Parallel between MD and CMI approaches for sequential backward elimination

MD	CMI
$\max\limits_{f_i \in S^t} I(S^{t+1}; C) =$	$\min\limits_{f_i \in S^t} I(\neg S^{t+1}; C \mid S^{t+1})$
$\max\limits_{f_i \in S^t} I(S^t \backslash f; C) =$	$\min\limits_{f_i \in S^t} I(\{\neg S^t, f_i\} \backslash f_i; C \mid S^t \backslash f_i)$
$\max\limits_{f_i \in S^t} I(S^t; C)^a + \max\limits_{f_i \in S^t}(-I(f_i; C \mid S^t \backslash f_i))$ $\Downarrow$ $\min\limits_{f_i \in S^t} I(f_i; C \mid S^t \backslash f_i)$	$\min\limits_{f_i \in S^t} I(\neg S^t; C \mid S^t)^b + \min\limits_{f_i \in S^t}(I(f_i; C \mid S^t \backslash f_i))$ $\Downarrow$ $\min\limits_{f_i \in S^t} I(f_i; C \mid S^t \backslash f_i)$

[a] This term is independent of f_i

[b] This term has the same value $\forall f_i$

The first term on the right-hand side of (26) measures the individual relevance of the candidate feature f_i with respect to output C; the second term measures the redundance of the candidate feature with the feature subset of previously selected features S^t; and the third term measures the complementarity between S^t and f_i in the context of C. However, from the practical point of view, Eq. (26) presents the difficulty of estimating MI in high-dimensional spaces, due to the presence of the set S^t in the second and third terms.

In what follows, we take a detour from the derivation of Brown *et al*. [10], using our own alternative approach. To avoid the previously mentioned problem, $I(f_i;S^t)$ with $|S^t| = p$ can be calculated by averaging all expansions over every single feature in S, by using the chain rule as follows:

$$\begin{array}{cccc} I(f_i;S^t) = & I(f_i;s_1) + & I(f_i;\neg s_1|s_1) \\ I(f_i;S^t) = & I(f_i;s_2) + & I(f_i;\neg s_2|s_2) \\ \vdots \quad = & \vdots & \vdots \\ I(f_i;S^t) = & I(f_i;s_p) + & I(f_i;\neg s_p|s_p) \\ \hline I(f_i;S^t) = & \multicolumn{2}{l}{\frac{1}{|S^t|}\sum\limits_{s_j\in S^t} I(f_i;s_j) + \frac{1}{|S^t|}\sum\limits_{s_j\in S^t} I(f_i;\neg s_j|s_j).} \end{array} \tag{27}$$

Analogously, we can obtain the following expansion for the conditional mutual information, $I(f_i;S^t|C)$:

$$I(f_i;S^t|C) = \frac{1}{|S^t|}\sum_{s_j\in S^t} I(f_i;s_j|C) + \frac{1}{|S^t|}\sum_{s_j\in S^t} I(f_i;\neg s_j|\{C,s_j\}). \tag{28}$$

Substituting (27) and (28) into Eq. (26) yields:

$$I(f_i;C|S^t) = I(f_i;C) - \left(\frac{1}{|S^t|}\sum_{s_j\in S^t} I(f_i;s_j) + \frac{1}{|S^t|}\sum_{s_j\in S^t} I(f_i;\neg s_j|s_j)\right) + \left(\frac{1}{|S|}\sum_{s_j\in S} I(f_i;s_j|C) + \frac{1}{|S|}\sum_{s_j\in S} I(f_i;\neg s_j|\{C,s_j\})\right). \tag{29}$$

Equation (29) can be approximated by considering assumptions of lower-order dependencies between features [3]. Features $s_j \in S^t$ are assumed to have only one-to-one dependencies with f_i or C. Formally, assuming statistical independence:

$$\begin{aligned} p(f_i|S^t) &= \prod_{s_j\in S^t} p(f_i|s_j) \\ p(f_i|\{S^t,C\}) &= \prod_{s_j\in S^t} p(f_i|\{s_j,C\}), \end{aligned} \tag{30}$$

we obtain the following low-order approximation:

$$I(f_i;C|S^t) \approx I(f_i;C) - \frac{1}{|S^t|}\sum_{s_j\in S^t} I(f_i;s_j) + \frac{1}{|S^t|}\sum_{s_j\in S^t} I(f_i;s_j|C). \tag{31}$$

Notice that Eq. (31) is an approximation of the multidimensional MI expressed by Eq. (26). Interestingly, Brown et al. [10] deduced a similar formula but with coefficients $1/|S^t|$ replaced by unity constants.

Equation (31) allows deriving some well-known heuristic feature selection methods. When only the first two terms of Eq. (31) are taken into account, it corresponds exactly to the minimal redundance maximal relevance (mRMR) criterion proposed in [46]. Moreover, if the term $1/|S|$ is replaced by a user-defined parameter β, then we obtain the MIFS criterion (*Mutual Information Feature Selection*) proposed by Battiti [4]. When considering only the first term in Eq. (31), we obtain the MIM criterion [39].

Equation (31) with its three terms corresponds exactly to the Joint Mutual Information (JMI)[10, 66]. Also, it corresponds with the Conditional Infomax Feature Extraction (CIFE) criterion proposed in [40] when the coefficient $|S^t| = 1$, $\forall t$. Moreover, the Conditional Mutual Information-based Feature Selection (CMIFS) criterion proposed in [12] is an approximation of Eq. (29), where only 0, 1, or 2 out of t summation terms are considered in each term. The CMIFS criterion is the following:

$$J_{cmifs}(f_i) = I(f_i;C) - I(f_i;s_t) + \sum_{s_j\in S; j\in\{1,t\}} I(f_i;s_j|C) - I(f_i;s_t|s_1). \tag{32}$$

The previously mentioned methods do not take into account the terms containing $\neg s_j$ in Eq. (29). This entails the assumption that f_i and $\neg s_j$ are independent, therefore $(I(f_i;\neg s_j) = I(f_i;\neg s_j|C) = 0)$. This approximation can generate errors in the sequential selection or backward elimination of variables. In order to somehow take into account the missing terms, let us consider the following alternative approximation of $I(f_i;C|S^t)$:

$$\begin{aligned} I(f_i;C|S^t) &= I(f_i;C) + I(f_i;S^t;C) \\ &= I(f_i;C) + I(f_i;\{s_j,\neg s_j\};C) \\ &= I(f_i;C) + I(f_i;s_j;C) + I(f_i;\neg s_j;C|s_j) \\ &= I(f_i;C|s_j) + I(f_i;\neg s_j;C|s_j). \end{aligned} \tag{33}$$

Averaging this decomposition over every single feature $s_j \in S^t$ we have:

$$I(f_i;C|S^t) = \frac{1}{|S^t|}\sum_{s_j\in S^t} I(f_i;C|s_j) + \frac{1}{|S^t|}\sum_{s_j\in S^t} I(f_i;\neg s_j;C|s_j). \tag{34}$$

The Interaction Capping (ICAP) [29] criterion approximates Eq. (33) by the following expression:

$$J_{\text{icap}}(f_i) = I(f_i;C) + \sum_{s_j\in S}\min(0, I(f_i;s_j;C))). \tag{35}$$

In ICAP [29], the information of variable f_i is penalized when the interaction between f_i, s_j, and C becomes redundant $(I(f_i;s_j;C) < 0)$, but the complementarity relationship among variables is neglected when $I(f_i;s_j;C) > 0$. The authors considered a Naive Bayes classifier, which assumes independence between variables.

Equation (34) allows deriving the Conditional Mutual Information Maximization (CMIM) criterion [21] when we consider only the first term on the right-hand side of this equation and replace the mean operator with a minimum operator. CMIM discards the second term on the right-hand side of Eq.(34) completely, taking into account only one-to-one relationships among variables and neglecting the multi-information among f_i, $\neg s_j$ and C in the context of s_j $\forall$ j. On the other hand, CMIM-2 [60] criterion corresponds exactly to the first term on the right-hand side of Eq. (34). These methods are able to detect pairs of relevant variables that act complementarily in predicting the class. In general, CMIM-2 outperformed CMIM in experiments using artificial and benchmark datasets [60].

So far we have reviewed feature selection approaches that avoid estimating MI in high-dimensional spaces. Bonev et al. [9] proposed an extension of the MD criterion, called Max-min-Dependence (MmD), which is defined as follows:

$$J_{\text{MmD}}(f_i) = I(\{f_i,S\};C) - I(\neg\{f_i,S\};C). \tag{36}$$

The procedure starts with the empty set $S = \emptyset$ and sequentially generates S^{t+1} as:

$$S^{t+1} = S^t \cup \max_{f_i\in F\setminus S}(J_{\text{MmD}}(f_i)). \tag{37}$$

The MmD criterion is heuristic and is not derived from a principled approach. However, Bonev et al. [9] were one of the first in selecting variables estimating MI in high-dimensional spaces [27], which allows using set of variables instead of individual variables. Chow and Huang [13] proposed combining a pruned Parzen window estimator with quadratic mutual information [47], using Renyi entropies, to estimate directly the MI between the feature subset S^t and the classes C, $I(S^t;C)$, in an effective and efficient way.

6 Open problems

In this section, we present some open problems and challenges in the field of feature selection, in particular from the point of view of information-theoretic methods. Here can be found a non-exhaustive list of open problems or challenges.

1. *Further developing a unifying framework for information-theoretic feature selection.* As we reviewed in Sect. 5, a unifying framework able to explain the advantages and limitations of successful heuristics has been proposed. This theoretical framework should be further developed in order to derive new efficient feature selection algorithms that include in their functional terms information related to the three types of features: relevant, redundant, and complementary. Also a stronger connection between this framework and the Markov blanket is needed. Developing hybrid methods that combine maximal dependency with minimal conditional mutual information is another possibility.
2. *Further improving the efficacy and efficiency of information-theoretic feature selection methods in high-dimensional spaces.* The computational time depends on the search strategy and the evaluation criterion [24]. As we enter the era of Big Data, there is an urgent need for developing very fast feature selection methods able to work with millions of features and billions of samples. An important challenge is developing more efficient methods for estimating MI in high-dimensional spaces. Automatically determining the optimal size of the feature subset is also of interest, many feature selection methods do not have a stop criterion. Developing new search strategies that go beyond greedy optimization is another interesting possibility.
3. *Further investigating the relationship between mutual information and Bayes error classification.* So far

lower and upper bounds for error classification have been found for the case of one random variable and the target class. Extending these results to the case of mutual information between feature subsets and the target class is an interesting open problem.

4. *Further investigating the effect of a finite sample over the statistical criteria employed and in MI estimation*. Guyon *et al*. [24] argued that feature subsets that are not sufficient may render better performance than sufficient feature subsets. For example, in the bioinformatics domain, it is common to have very large input dimensionality and small sample size [49].
5. *Further developing a framework for studying the relationship between feature selection and causal discovery*. Guyon et al. [25] investigated causal feature selection. The authors argued that the knowledge of causal relationships can benefit feature selection and viceversa. A challenge is to develop efficient Markov blanket induction algorithms for non-faithful distributions.
6. *Developing new criteria of statistical dependence beyond correlation and MI*. Seth and Principe [51] revised the postulates of measuring dependence according to Renyi, in the context of feature selection. An important topic is normalization, because a measure of dependence defined on different kinds of random variables should be comparable. There is no standard theory about MI normalization [16, 19]. Another problem is that estimators of measures of dependence should be good enough, even when using a few realizations, in the sense of following the desired properties of these measures. Seth and Principe [51] argued that this property is not satisfied by MI estimators, because they do not reach the maximum value under strict dependence, and are not invariant to one-to-one transformations.

7 Conclusions

We have presented a review of the state of the art in information-theoretic feature selection methods. We showed that modern feature selection methods must go beyond the concepts of relevance and redundance to include complementarity (synergy). In particular, new feature selection methods that assess features in context are necessary. Recently, a unifying framework has been proposed, which is able to retrofit successful heuristic criteria. In this work, we have further developed this framework, presenting some new results and derivations. The unifying theoretical framework allows us to indicate the approximations made by each method and therefore their limitations. A number of open problems in the field are suggested as challenges for the avid reader.

Acknowledgments This work was funded by CONICYT-CHILE under grant FONDECYT 1110701.

References

1. Almuallim H, Dietterich TG (1991) Learning with many irrelevant features. In: Artificial intelligence, proceedings of the ninth national conference on, AAAI Press, pp 547–552
2. Almuallim H, Dietterich TG (1992) Efficient algorithms for identifying relevant features. In: Artificial intelligence, proceedings of the ninth canadian conference on, Morgan Kaufmann, pp 38–45
3. Balagani K, Phoha V (2010) On the feature selection criterion based on an approximation of multidimensional mutual information. IEEE Trans Pattern Anal Mach Intell 32(7):1342–1343
4. Battiti R (1994) Using mutual information for selecting features in supervised neural net learning. IEEE Trans Neural Netw 5(4):537–550
5. Bell AJ (2003) The co-information lattice. Analysis pp 921–926
6. Bell DA, Wang H (2000) A formalism for relevance and its application in feature subset selection. Mach Learn 41(2):175–195
7. Bellman RE (1961) Adaptive control processes: a guided tour. 1st edn. Princeton University Press, Princeton
8. Bins J, Draper B (2001) Feature selection from huge feature sets. In: Computer Vision, 2001. Proceedings eighth IEEE international conference, vol 2, pp 159–165
9. Bonev B, Escolano F, Cazorla M (2008) Feature selection, mutual information, and the classification of high-dimensional patterns: applications to image classification and microarray data analysis. Pattern Anal Appl 11(3-4):309–319
10. Brown G, Pocock A, Zhao MJ, Luján M (2012) Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. J Mach Learn Res 13:27–66
11. Brown LE, Tsamardinos I (2008) Markov blanket-based variable selection in feature space. Technical report dsl-08-01, Discovery systems laboratory, Vanderbilt University
12. Cheng H, Qin Z, Feng C, Wang Y, Li F (2011) Conditional mutual information-based feature selection analyzing for synergy and redundancy. Electron Telecommun Res Inst 33(2):210–218
13. Chow T, Huang D (2005) Estimating optimal feature subsets using efficient estimation of high-dimensional mutual information. IEEE Trans Neural Netw 16(1):213–224
14. Cover TM, Thomas JA (2006) Elements of Information Theory. 2nd edn. Wiley-Interscience, New Jersey
15. Davies S, Russell S (1994) Np-completeness of searches for smallest possible feature sets. In: Intelligent Relevance, association for the advancement of artificial intelligence symposium on, AAAI Press, pp 37–39
16. Duch W (2006) Filter methods. In: Feature extraction, foundations and applications, studies in fuzziness and soft computing, vol 207, Springer, Heidelberg, chap 3, pp 167–185
17. Duch W, Winiarski T, Biesiada J, Kachel A (2003) Feature selection and ranking filter. In: International conference on artificial neural networks (ICANN) and International conference on neural information processing (ICONIP), pp 251–254
18. Estévez PA, Caballero R (1998) A niching genetic algorithm for selecting features for neural networks classifiers. In: Perspectives in neural computation, Springer, New York, pp 311–316
19. Estévez PA, Tesmer M, Pérez CA, Zurada JM (2009) Normalized mutual information feature selection. IEEE Trans Neural Netw 20(2):189–201
20. Feder M, Merhav N (1994) Relations between entropy and error probability. IEEE Trans Inform Theory 40(1):259–266
21. Fleuret F, Guyon I (2004) Fast binary feature selection with conditional mutual information. J Mach Learn Res 5:1531–1555

22. Guo B, Nixon MS (2009) Gait feature subset selection by mutual information. Syst Man Cybern Part A IEEE Trans 39(1):36–46
23. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
24. Guyon I, Elisseeff A (2006) An introduction to feature extraction. In: Feature extraction, foundations and applications, studies in fuzziness and soft computing, vol 207, Springer, Berlin, pp 1–25
25. Guyon I, Aliferis C, Elisseeff A (2008) Causal feature selection. In: Liu H, Motoda H (eds) Computational methods of feature selection, Chapman & Hall/CRC, chap 4
26. Hellman M, Raviv J (1970) Probability of error, equivocation, and the chernoff bound. IEEE Trans Inform Theory 16(4):368–372
27. Hero A, Michel O (1999) Estimation of renyi information divergence via pruned minimal spanning trees. Higher-Order statistics proceedings of the IEEE signal processing workshop, pp 264–268
28. Huang S (2003) Dimensionality reduction in automatic knowledge acquisition: a simple greedy search approach. IEEE Trans Knowl Data Eng 15(6):1364–1373
29. Jakulin A (2005) Learning based on attribute interactions. PhD thesis, University of Ljubljana, Slovenia
30. Jakulin A, Bratko I (2003) Quantifying and visualizing attribute interactions. CoRR cs.AI/0308002, http://arxiv.org/abs/cs.AI/0308002
31. Kohavi R, John GH (1997) Wrappers for feature subset selection. Artif Intell 97(1–2):273–324
32. Kojadinovic I (2005) Relevance measures for subset variable selection in regression problems based on k-additive mutual information. Comput Stat Data Anal 49(4):1205–1227
33. Koller D, Sahami M (1996) Toward optimal feature selection. Technical Report 1996-77, Stanford InfoLab
34. Kraskov A, Stögbauer H, Grassberger P (2004) Estimating mutual information. Phys Rev E 69:066,138
35. Kullback S (1997) Information theory and statistics. 2nd edn. Dover, New York
36. Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22:49–86
37. Kwak N, Choi CH (2002) Input feature selection for classification problems. IEEE Trans Neural Netw 13(1):143–159
38. Lal KN, Chapelle O, Weston J, Elisseeff A (2006) Embedded methods. In: Feature extraction, foundations and applications, studies in fuzziness and soft computing, vol 207, Springer, Berlin, chap 5, pp 167–185
39. Lewis DD (1992) Feature selection and feature extraction for text categorization. In: Proceedings of speech and natural language workshop, Morgan Kaufmann, pp 212–217
40. Lin D, Tang X (2006) Conditional infomax learning: an integrated framework for feature extraction and fusion. In: Computer vision—ECCV 2006, Lecture Notes in Computer Science, vol 3951, Springer, Berlin, pp 68–82
41. Marill T, Green D (1963) On the effectiveness of receptors in recognition systems. IEEE Trans Inform Theory 9(1):11–17
42. McGill W (1954) Multivariate information transmission. Psychometrika 19(2):97–116
43. Meyer P, Schretter C, Bontempi G (2008) Information-theoretic feature selection in microarray data using variable complementarity. IEEE J Sel Top Signal Process 2(3):261–274
44. Mo D, Huang SH (2011) Feature selection based on inference correlation. Intell Data Anal 15(3):375–398
45. Mo D, Huang SH (2012) Fractal-based intrinsic dimension estimation and its application in dimensionality reduction. IEEE Trans Knowl Data Eng 24(1):59–71
46. Peng H, Long F, Ding C (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27(8):1226–1238
47. Principe JC (2010) Information theoretic learning: Renyi's entropy and kernel perspectives. 1st edn. Springer Publishing Company, Berlin
48. Raudys S, Jain A (1991) Small sample size effects in statistical pattern recognition: recommendations for practitioners. IEEE Trans Pattern Anal Mach Intell 13(3):252–264
49. Saeys Y, Inza I, Larranaga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23(19):2507–2517
50. Sebban M, Nock R (2002) A hybrid filter/wrapper approach of feature selection using information theory. Pattern Recogn 35(4): 835–846
51. Seth S, Príncipe J (2010) Variable selection: A statistical dependence perspective. In: Machine Learning and Applications (ICMLA), 2010 ninth international conference, pp 931–936
52. Shannon CE (1948) A mathematical theory of communication. Bell System Technical J 27:379–423, 625–56
53. Somol P, Pudil P, Kittler J (2004) Fast branch & bound algorithms for optimal feature selection. IEEE Trans Pattern Anal Mach Intell 26(7):900–912
54. Stearns SD (1976) On selecting features for pattern classifiers. In: Pattern recognition, proceedings of the 3rd international conference on, Coronado, CA, pp 71–75
55. Tishby N, Pereira FC, Bialek W (1999) The information bottleneck method. Proceedings on 37th Annu Allerton conference communication, Control and Computing
56. Torkkola K (2006) Information-theoretic methods. In: Feature extraction, foundations and applications, studies in fuzziness and soft computing, vol 207, Springer, Berlin, chap 6, pp 167–185
57. Trunk GV (1979) A problem of dimensionality: A simple example. IEEE Trans Pattern Anal Mach Intell-1(3):306–307
58. Tsamardinos I, Aliferis CF (2003) Towards principled feature selection: relevancy, filters and wrappers. In: Artificial intelligence and statistics, proceedings of the ninth international workshop, Morgan Kaufmann Publishers
59. Tsamardinos I, Aliferis CF, Statnikov E (2003) Algorithms for large scale markov blanket discovery. In: The 16th international FLAIRS conference, St, AAAI Press, pp 376–380
60. Vergara JR, Estévez PA (2010) Cmim-2: an enhanced conditional mutual information maximization criterion for feature selection. J Appl Comput Sci Methods 2(1):5–20
61. Vidal-Naquet M, Ullman S (2003) Object recognition with informative features and linear classification. In: Computer Vision, 2003. Proceedings ninth IEEE international conference, vol 1, pp 281–288
62. Watanabe S (1960) Information theoretical analysis of multivariate correlation. IBM J Res Dev 4(1):66–82
63. Webb AR (2002) Statistical Pattern Recognition. 2nd edn. Wiley, New Jersey
64. Weston J, Mukherjee S, Chapelle O, Pontil M, Poggio T, Vapnik V (2000) Feature selection for svms. In: Advances in neural information processing systems 13, MIT Press, pp 668–674
65. Whitney A (1971) A direct method of nonparametric measurement selection. IEEE Trans Comput C-20(9):1100–1103
66. Yang HH, Moody J (1999) Feature selection based on joint mutual information. In: Advances in intelligent data analysis, proceedings of international ICSC symposium, pp 22–25
67. Yu L, Liu H (2004) Efficient feature selection via analysis of relevance and redundancy. J Mach Learn Res 5:1205–1224
68. Zhao Z, Liu H (2009) Searching for interacting features in subset selection. Intell Data Anal 13(2):207–228