Pattern Recognition 53 (2016) 46–58

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Can high-order dependencies improve mutual information based feature selection?

Nguyen Xuan Vinh *,[1], Shuo Zhou, Jeffrey Chan, James Bailey

Department of Computing and Information Systems, The University of Melbourne, Victoria, Australia

ARTICLE INFO

Article history:
Received 10 March 2015
Received in revised form
7 September 2015
Accepted 10 November 2015
Available online 19 November 2015

Keywords:
Feature selection
Mutual information
High-order dependency

ABSTRACT

Mutual information (MI) based approaches are a popular paradigm for feature selection. Most previous methods have made use of low-dimensional MI quantities that are only effective at detecting low-order dependencies between variables. Several works have considered the use of higher dimensional mutual information, but the theoretical underpinning of these approaches is not yet comprehensive. To fill this gap, in this paper, we systematically investigate the issues of employing high-order dependencies for mutual information based feature selection. We first identify a set of assumptions under which the original high-dimensional mutual information based criterion can be decomposed into a set of low-dimensional MI quantities. By relaxing these assumptions, we arrive at a principled approach for constructing higher dimensional MI based feature selection methods that takes into account higher order feature interactions. Our extensive experimental evaluation on real data sets provides concrete evidence that methodological inclusion of high-order dependencies improve MI based feature selection.

1. Introduction

Feature selection is an important task in data mining and knowledge discovery. Effective feature selection can improve performance while reducing the computational cost of learning systems. In this paper, we focus on mutual information (MI) based feature selection, which is a very popular *filter* paradigm. Compared to *wrapper* and *embedded* approaches [1], filter methods, such as those based on the MI criteria, are generally less optimized, but possess the major advantage of being learning-model independent and also typically less computationally intensive.

MI based feature selection is concerned with identifying a subset $\mathbf{S}$ of m features $\{X_1,\ldots,X_m\}$ within the original set $\mathbf{X}$ of M features in a data set, that maximizes the multidimensional joint MI between features and the class variable C, defined as

$$I(\mathbf{S};C) \triangleq \sum_{X_1,\ldots,X_m,C} P(X_1,\ldots,X_m,C)\log\frac{P(X_1,\ldots,X_m,C)}{P(X_1,\ldots,X_m)P(C)} \quad (1)$$

This criterion possesses a solid theoretical foundation, in that the MI can be used to write both an upper and lower bound on the Bayes error rate [2,3]. Nevertheless, the problems of estimating high-dimensional joint MI, and more generally estimating high-dimensional probability distribution, especially from small samples, are long-standing challenges in statistics. Therefore, a rich body of work in the MI-based feature selection literature approaches this difficulty by approximating the high-dimensional joint MI with low-dimensional MI terms. A particularly popular and successful class of methods makes use of the following criterion, which is the combination of low-dimensional MI terms known as '*relevancy*' and '*redundancy*',

$$f(X_m) \triangleq I(X_m;C) - \beta\sum_{X_j\in\mathbf{S}} I(X_m;X_j) \quad (2)$$

Under this framework, the features are often selected in an incremental manner: given a set $\mathbf{S}$ of $m-1$ already selected features $\{X_1,\ldots,X_{m-1}\}$, the next feature X_m is selected so that $f(X_m)$ is maximized. The term $I(X_m;C)$ measures the relevancy of X_m to the class variable C, while $\sum_{X_j\in\mathbf{S}} I(X_m;X_j)$ quantifies the redundancy between X_m and the selected features in $\mathbf{S}$, and β plays the role of a balancing factor. Many MI-based feature selection heuristics can be shown to be variations of (2) [3], including highly influential methods such as the Mutual Information Feature Selection (MIFS) criterion ($\beta\in[0,1]$) [4], and the Minimum Redundancy Maximum Relevance (MRMR) criterion ($\beta=1/|\mathbf{S}|$) [5].

It is noted that the two-dimensional MI can only detect pairwise variable interactions, either between two features or between a feature and the class variable. More complicated variable interactions cannot be identified with the two-dimensional MI. Fig. 1 provides an illustrative example of two variables (switches) that

* Corresponding author. Tel.: +614 3220 8948.
E-mail address: vinh.nguyen@unimelb.edu.au (N.X. Vinh).
[1] Postal address: Department of Computing and Information Systems, The University of Melbourne, Parkville VIC 3010, Australia.

http://dx.doi.org/10.1016/j.patcog.2015.11.007
0031-3203/

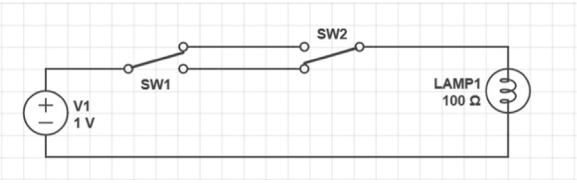

Fig. 1. An example of high-order variable interaction.

jointly control the target variable (the lamp). Knowing the state of either switch alone provides no information about whether the lamp is on or off. Only the joint state of both switches provides comprehensive knowledge on the state of the lamp. The pairwise mutual information cannot detect this type of multi-variable interaction.

To address this shortcoming, several works have considered the use of higher-dimensional MI quantities, such as the joint relevancy $I(X_iX_j;C)$ [6], the conditional relevancy $I(X_i;C|X_j)$ [3] and the conditional redundancy $I(X_i;X_j|C)$ [7]. Brown et al. [3] showed that many such proposed methods can fit within the parameterized criterion:

$$J(X_m) \triangleq I(X_m;C) - \beta \sum_{X_j \in \mathbf{S}} I(X_m;X_j) + \gamma \sum_{X_j \in \mathbf{S}} I(X_m;X_j|C). \tag{3}$$

For example, the Joint Mutual Information (JMI) criterion [6] can be obtained with $\beta = \gamma = 1/|\mathbf{S}|$. The Conditional Informative Feature Extraction (CIFE) criterion [8] is obtained with $\beta = \gamma = 1$. The extended MRMR criterion [9] is a special case when $\beta = \gamma$. The objective in (2), including MRMR and MIFS, are clearly special cases where $\gamma = 0$. These methods can detect higher order variable dependencies, in particular those between two features and the class variable. However, all the mentioned criteria were handcrafted and their theoretical underpinning is not well understood. In particular, (i) *in retrospect*, we would like to understand how these criteria are related to the original full joint MI criterion in (1), and (ii) *moving forward*, we would like to leverage this understanding to design higher-order MI based feature selection methods in a more systematic and methodological manner. Recent work has partially elucidated the former question [10,3], while to our knowledge, the latter question has not been investigated.

Contributions: To address the identified gap, in this paper, we study the connection between the low-dimensional MI based criteria, such as the ones in (2) and (3), and the ultimate high-dimensional MI objective in (1). The benefit of such an investigation is two-fold: (i) to establish the theoretical underpinnings for heuristics based on (2) and (3), and (ii) to inspire a systematic and methodological development of higher-dimensional MI-based feature selection techniques by relaxing the identified assumptions. We take a first step towards this direction by proposing several novel MI based feature selection approaches that take into account higher-order dependency between features, in particular three-way feature interaction $I(X_i;X_j|X_k)$. Our extensive experimental evaluation shows that systematic inclusion of higher-dimensional MI quantities improves the feature selection performance.

2. Assumptions underlying low-dimensional MI-based feature selection heuristics

Our first goal in this paper is to strive for a more comprehensive understanding of the theoretical underpinnings behind various MI based feature selection heuristics. Several recent works have partially addressed this question. Balagani and Proha [10] identified a set of assumptions underlying the objective (2) while Brown et al. [3] investigated the assumptions underlying the more general objective (3). In this section, we continue to develop further along these lines, while making some new connections between the previous work.

In [10], Balagani and Proha set out to identify the conditions under which the high-dimensional MI in (1) could be decomposed exactly as a sum of low-dimensional relevancy and redundancy MI terms, i.e.,

$$I(\mathbf{S};C) \equiv \sum_{i=1}^{m} I(X_i;C) - \sum_{i=2}^{m} \sum_{j<i} I(X_i;X_j) \tag{4}$$

They showed that under the following three assumptions, the identity (4) holds true.

Assumption 1. The selected features $\{X_1, X_2, \ldots, X_{m-1}\}$ are independent, i.e.,

$$P(X_1, X_2, \ldots, X_{m-1}) = \prod_{i=1}^{m-1} P(X_i) \tag{5}$$

Assumption 2. The selected features $\{X_1, X_2, \ldots, X_{m-1}\}$ are conditionally independent given the feature X_m, i.e.,

$$P(X_1, X_2, \ldots, X_{m-1}|X_m) = \prod_{i=1}^{m-1} P(X_i|X_m). \tag{6}$$

Assumption 3 (*Naive Bayes independence assumption*). Each feature independently influences the class variable, i.e.,

$$P(X_m|X_1, \ldots, X_{m-1}, C) = P(X_m|C). \tag{7}$$

We will argue here briefly that, of these three assumptions, Assumption 1 is a strong condition. More specifically, the condition in (5) implies that all features in $\mathbf{S}$ are pairwise independent, indeed

$$\begin{aligned} \forall X_i, X_j \in \mathbf{S} : P(X_i, X_j) &= \sum_{\mathbf{S}\setminus\{X_i,X_j\}} P(X_1, X_2, \ldots, X_{m-1}) \\ &= \sum_{\mathbf{S}\setminus\{X_i,X_j\}} P(X_1)P(X_2)\ldots P(X_{m-1}) = P(X_i)P(X_j) \end{aligned}$$

Furthermore, since at design time, it is not possible to anticipate which features of $\mathbf{X}$ will be selected in $\mathbf{S}$, it is necessary that all features in the original feature set $\mathbf{X}$ are also pairwise independent, for the identity (4) to hold true on any selected subset of $\mathbf{X}$. Therefore, with this assumption, we effectively have $I(X_i;X_j) = 0 \ \forall i \neq j$, implying that the incremental objective in (2) reduces to the simplistic objective of $f(X_m) = I(X_m;C)$, i.e., selecting the m-th highest ranking feature, in terms of the MI shared with C, without taking into account the redundancy with the selected features.

2.1. An alternative view

In this section, we present an alternative view on the issue of approximating high-dimensional MI with low-dimensional MI terms. First, note that even if the high-dimensional MI were easily estimable, the problem of identifying a subset $\mathbf{S}$ that shares the maximal MI with C remains a challenging combinatorial optimization problem without known efficient solution. An exhaustive search will be of $O(2^M)$ time complexity, while restricting the maximum size of $\mathbf{S}$ to $k < M$ will reduce the cost to $O(M^k)$, but will still be expensive. As such, an obvious iterative greedy strategy is to select one feature at a time: given the set $\mathbf{S} = \{X_1, \ldots, X_{m-1}\}$ of $m-1$ already selected features, the m-th feature is chosen maximizing the following objective function:

$$\arg\max_{X_m \in \mathbf{X}\setminus\mathbf{S}} I(\mathbf{S} \cup X_m; C) \tag{8}$$

We will now try to understand under what conditions, low-order MI based heuristics such as MRMR and MIFS in (2) will produce

the same result as (8), i.e.,

$$\arg\max_{X_m \in \mathbf{X}\setminus\mathbf{S}} I(\mathbf{S}\cup X_m; C) \equiv \arg\max_{X_m \in \mathbf{X}\setminus\mathbf{S}} I(X_m; C) - \sum_{X_j \in \mathbf{S}} I(X_m; X_j) \tag{9}$$

Comparing (9) and (4), there is a subtle yet critical difference between our viewpoint and Balagani and Proha's: while Balagani and Proha aim to match the global objective function, we aim at matching the outcome of the greedy iterative optimization procedure. We point out here that MRMR [4] and MIFS [5], among other similar heuristics, aim to approximate the incremental optimization problem in (8), rather than to approximate the original joint mutual information criterion $I(\mathbf{S}; C)$. We now prove the following result.

Theorem 1. *Under Assumptions 2 and 3, the equality (9) holds true.*

Proof. From the chain rule of mutual information, we have $I(\mathbf{S}\cup X_m; C) = I(\mathbf{S}; C) + I(X_m; C|\mathbf{S})$. Since $I(\mathbf{S}; C)$ remains constant w.r.t. X_m, we have $\text{argmax}_{X_m \in \mathbf{X}\setminus\mathbf{S}} I(\mathbf{S}\cup X_m; C) \equiv \text{argmax}_{X_m \in \mathbf{X}\setminus\mathbf{S}} I(X_m; C|\mathbf{S})$. As proven in [11], the conditional MI $I(X_m; C|\mathbf{S})$ can be expressed as

$$I(X_m; C|\mathbf{S}) = I(X_m; C) - [I(X_m; \mathbf{S}) - I(X_m; \mathbf{S}|C)], \tag{10}$$

we can therefore match the 'relevancy' $I(X_m; C)$ term. Next, we need to match the 'redundancy' term, i.e.,

$$\arg\min_{X_m \in \mathbf{X}\setminus\mathbf{S}} I(X_m; \mathbf{S}) - I(X_m; \mathbf{S}|C) \equiv \arg\min_{X_m \in \mathbf{X}\setminus\mathbf{S}} \sum_{X_j \in \mathbf{S}} I(X_m; X_j) \tag{11}$$

It is easily seen that under Assumption 3, $I(X_m; \mathbf{S}|C) = H(X_m|C) - H(X_m|C, \mathbf{S}) = H(X_m|C) - H(X_m|C) = 0$. Further, under Assumption 2

$$\begin{aligned} I(X_m; \mathbf{S}) &= H(\mathbf{S}) - H(\mathbf{S}|X_m) = H(\mathbf{S}) - \sum_{X_j \in \mathbf{S}} H(X_j|X_m) \\ &= H(\mathbf{S}) - \sum_{X_j \in \mathbf{S}} H(X_j) + \sum_{X_j \in \mathbf{S}} I(X_j; X_m) \end{aligned} \tag{12}$$

Taking into account the fact that $H(\mathbf{S}) - \sum_{X_j \in \mathbf{S}} H(X_j)$ is constant w.r.t. X_m, we have that (11) holds true.□

Thus, it can be seen that by matching the outcome of the actual incremental optimization procedure, but not the objective function, we are now able to drop the strong Assumption 1. Note that for this theoretical analysis, we have omitted the balancing factor β, which is of a heuristical nature. β was originally introduced in MIFS [4] and MRMR [5] to balance the relevancy and redundancy terms. If Assumptions 2 and 3 hold true, then naturally β is unnecessary, as all the equalities hold in an exact sense. Therefore, β can be regarded as a practical adjustment to be used when the required assumptions do not hold.

2.2. An alternative sufficient condition set

The decomposition (10) of the conditional MI $I(X_m; C|\mathbf{S})$, as observed in [11], brings about an interesting insight: the 'total redundancy' comprises an unconditional redundancy term $I(X_m; \mathbf{S})$, minus a class-conditional redundancy term $I(X_m; \mathbf{S}|C)$. In MIFS/MRMR formulation in (2), only the unconditional redundancy was considered. This is a result of Assumption 3, under which $I(X_m; \mathbf{S}|C)$ vanishes, while $I(X_m; \mathbf{S})$ is decomposed into a sum of pairwise MI terms under Assumption 2. In this section, we investigate the matter further by asking, provided we do not use the naive Bayes independence Assumption 3, what other assumption is needed to decompose $I(X_m; \mathbf{S}|C)$ into sums of low-dimensional MI terms. Brown et al. [3] proposed such an assumption, which can be seen as an analogue to Assumption 2, as follows:

Assumption 3a. The selected features $\{X_1, X_2, \ldots, X_{m-1}\}$ are conditionally independent given the feature X_m and the class C, i.e.,

$$P(X_1, X_2, \ldots, X_{m-1}|C, X_m) = \prod_{i=1}^{m-1} P(X_i|C, X_m). \tag{13}$$

Now under Assumption 3a,

$$\begin{aligned} I(X_m; \mathbf{S}|C) &= H(\mathbf{S}|C) - H(\mathbf{S}|C, X_m) \\ &= H(\mathbf{S}|C) - \sum_{X_j \in \mathbf{S}} H(X_j|C, X_m) \\ &= H(\mathbf{S}|C) - \sum_{X_j \in \mathbf{S}} H(X_j|C) + \sum_{X_j \in \mathbf{S}} I(X_j; X_m|C) \end{aligned} \tag{14}$$

Substituting (14) into the l.h.s. of (11), and taking into account the fact that $H(\mathbf{S}|C) - \sum_{X_j \in \mathbf{S}} H(X_j|C)$ is constant w.r.t. X_m, then the problem of minimizing the 'total redundancy' is equivalent to

$$\arg\min_{X_m \in \mathbf{X}\setminus\mathbf{S}} I(X_m; \mathbf{S}) - I(X_m; \mathbf{S}|C) \equiv \arg\min_{X_m \in \mathbf{X}\setminus\mathbf{S}} \sum_{X_j \in \mathbf{S}} \{I(X_m; X_j) - I(X_m; X_j|C)\} \tag{15}$$

The new redundancy criterion in the r.h.s. of (15) is interesting, as it reflects closely the fact that the original high-dimensional redundancy term consists of an unconditional part, and a class-conditioned part (2nd and 3rd term of (10) respectively). Now, if we introduce an additional assumption:

Assumption 3b. Features in $\mathbf{S}$ and X_m are pairwise class-conditionally independent, i.e.,

$$P(X_m, X_j|C) = P(X_m|C)P(X_j|C) \;\; \forall X_j \in \mathbf{S}. \tag{16}$$

then it is easily seen that the class-conditioned redundancy terms $I(X_m; X_j|C)$'s in (15) will also vanish, and so (15) again reduces to (11). Thus together, Assumptions 2, 3a and 3b achieve the same effect as Assumptions 2 and 3. An interesting remark to note is that Assumption 3 is a strong condition, which can be proven to entail both Assumptions 3a and 3b as corollaries.

Theorem 2. *Assumption 3 implies Assumptions 3a and 3b as corollaries.*

Proof. (i) Assumption 3 ⇒ Assumption 3a: we factor the l.h.s. of (13) as

$$\begin{aligned} P(X_1, X_2, \ldots, X_{m-1}|C, X_m) &= P(X_1|C, X_m) \times P(X_2|C, X_m, X_1) \\ &\times \cdots \times P(X_{m-1}|C, X_m, X_1, \ldots, X_{m-2}) \end{aligned} \tag{17}$$

From Assumption 3 we have

$$\begin{aligned} P(X_2|C, X_m, X_1) &= P(X_2|C) \\ P(X_2|C, X_m) &= P(X_2|C) \end{aligned} \tag{18}$$

Thus $P(X_2|C, X_m, X_1) = P(X_2|C, X_m)$. Similarly,

$$P(X_3|C, X_m, X_1, X_2) = P(X_3|C, X_m) \tag{19}$$

$$\vdots$$

$$P(X_{m-1}|C, X_m, X_1, \ldots, X_{m-2}) = P(X_{m-1}|C, X_m) \tag{20}$$

Substituting into (17) we have

$$P(X_1, X_2, \ldots, X_{m-1}|C, X_m) = \prod_{i=1}^{m-1} P(X_i|C, X_m). \tag{21}$$

(ii) Assumption 3 ⇒ Assumption 3b: We factor the l.h.s. of (16) as

$$P(X_m, X_j|C) = P(X_j|C)P(X_m|X_j, C) = P(X_j|C)P(X_m|C) \tag{22}$$

with the last equality being due to $P(X_m|X_j, C) = P(X_m|C)$, as per Assumption 3.□

The advantage of adopting Assumptions 3a and 3b over Assumption 3 is that, besides making a set of weaker assumptions,

we can individually omit Assumption 3b, giving rise to a new class of heuristics that makes use of the class-conditioned redundancy, which is the objective in (3).

3. Relaxing the assumptions

In the previous section, we have studied the assumptions underlying low-dimensional MI-based criteria for feature selection. While the previous work [3,10] retrospectively investigated these assumptions in regards to existing heuristics, we go one step forward in asking how these assumptions can guide the systematic and methodological development of new approaches for MI based feature selection. First, recall from previous sections that our goal is to carry out $\arg\max_{X_m \in \mathbf{X}\setminus\mathbf{S}} I(\mathbf{S}\cup X_m; C)$ in an incremental fashion, and further recall that

$$\arg\max_{X_m \in \mathbf{X}\setminus\mathbf{S}} I(\mathbf{S}\cup X_m; C) \equiv \arg\max_{X_m \in \mathbf{X}\setminus\mathbf{S}} I(X_m; C\,|\,\mathbf{S}) \equiv \arg\max_{X_m \in \mathbf{X}\setminus\mathbf{S}} I(X_m; C) - [I(X_m;\mathbf{S}) - I(X_m;\mathbf{S}|\,C)] \tag{23}$$

We now pay attention to the high-dimensional redundancy term $I(X_m;\mathbf{S})$. Note that Assumption 2, which is needed for decomposing the high-dimensional redundancy term $I(X_m;\mathbf{S})$, can be relaxed to reflect the higher-order dependency between features. For example:

Assumption 2'. The selected features $\{X_1, X_2, \ldots, X_{m-1}\}$ are conditionally independent given the feature X_m and any feature $X_j \in \mathbf{S}$, i.e.,

$$P(X_1, X_2, \ldots, X_{m-1}\,|\,X_m) = P(X_j\,|\,X_m)\prod_{\substack{i=1\\ i\neq j}}^{m-1} P(X_i\,|\,X_m, X_j) \tag{24}$$

Under this relaxed assumption, we can show that

Theorem 3. *Under Assumption 2′ we have*

$$I(X_m;\mathbf{S}) = I(X_m;X_j) + \sum_{X_i\in\mathbf{S}; i\neq j} I(X_m;X_i\,|\,X_j) + \Omega \tag{25}$$

where Ω is a constant w.r.t. X_m.

Proof.

$$\begin{aligned} I(X_m;\mathbf{S}) &= H(\mathbf{S}) - H(\mathbf{S}|\,X_m) \\ &= H(\mathbf{S}) - \left\{ H(X_j\,|\,X_m) + \sum_{i=1, i\neq j}^{m-1} H(X_i\,|\,X_m, X_j) \right\} \\ &= H(\mathbf{S}) - H(X_j) + I(X_m;X_j) - \sum_{X_i\in\mathbf{S}; i\neq j} H(X_i\,|\,X_j) \\ &\quad + \sum_{X_i\in\mathbf{S}; i\neq j} I(X_i;X_m\,|\,X_j) = I(X_m;X_j) + \sum_{X_i\in\mathbf{S}; i\neq j} I(X_m;X_i\,|\,X_j) + \Omega \end{aligned}$$

where $\Omega = H(\mathbf{S}) - H(X_j) - \sum_{X_i\in\mathbf{S}; i\neq j} H(X_i\,|\,X_j)$ is constant w.r.t. X_m.□

To avoid the need of identifying a particular feature $X_j \in \mathbf{S}$ to condition on, this process can be averaged over all $X_j \in \mathbf{S}$, resulting in

$$I(X_m;\mathbf{S}) = \frac{1}{|\mathbf{S}|}\sum_{X_j\in\mathbf{S}} \left\{ I(X_m;X_j) + \sum_{\substack{X_i\in\mathbf{S}\\ i\neq j}} I(X_m;X_i\,|\,X_j) \right\} + \Omega' \tag{26}$$

where Ω' is also a constant w.r.t. X_m. This newly obtained redundancy quantity takes into account the second-order interactions between the features, i.e., the three-way feature interaction terms $I(X_m;X_i\,|\,X_j)$. We note that this is only one example of how our analysis in this paper could be useful to guide the systematic development of novel MI-based feature selection techniques that make use of higher-dimensional MI, e.g., 3-dimensional or higher, provided that the sample size is sufficiently large to allow reasonably accurate estimates. Assumptions 2' and 3a can be relaxed further in a similar manner to capture higher-order feature–feature and feature–class dependencies.

4. RelaxMRMR: a novel higher-order MI-based feature selection approach

In this section, we design a novel MI-based criterion for feature selection based on the theoretical analysis in Section 3. We shall make use of Assumptions 2' and 3a. By substituting the new redundancy measure in (26) into the objective in (23), we arrive at the following criterion, which is *exactly equivalent* to the high-dimensional MI objective $I(X_m; C\,|\,\mathbf{S})$:

Form − 0:

$$\max_{X_m\in\mathbf{X}\setminus\mathbf{S}} \left\{ I(X_m;C) - \frac{1}{|\mathbf{S}|}\sum_{X_j\in\mathbf{S}} \left\{ I(X_m;X_j) + \sum_{\substack{X_i\in\mathbf{S}\\ i\neq j}} I(X_m;X_i\,|\,X_j) \right\} + \sum_{X_j\in\mathbf{S}} I(X_m;X_j\,|\,C) \right\} \tag{27}$$

Unfortunately, in practice, these assumptions do not usually hold true. Therefore, some normalization is needed to get the right balance between different MI quantities, i.e., relevancy $I(X_m;C)$, redundancy $I(X_m;X_j)$, class-relevant redundancy $I(X_m;X_j\,|\,C)$ and second-order interaction $I(X_m;X_i\,|\,X_j)$. This normalization is similarly required by other successful heuristics, such as MRMR and JMI. In the ideal form of MRMR, there is also no need to regulate the weight between the relevance and redundancy. However, in reality, normalization is usually desired as the required assumptions for these criteria may not always hold true. In fact, according to Brown et al. [3], normalizing the redundancy terms by the selected feature set size is essential for a good criterion. This ensures that the relevancy of a feature remains informative when the number of selected features increases.

Our first attempt is to only normalize the class-relevant redundancy terms $I(X_m;X_j\,|\,C)$ by the number of selected features, resulting in:

Form − 1:

$$\max_{X_m\in\mathbf{X}\setminus\mathbf{S}} \left\{ I(X_m;C) - \frac{1}{|\mathbf{S}|}\sum_{X_j\in\mathbf{S}} I(X_m;X_j) + \frac{1}{|\mathbf{S}|}\sum_{X_j\in\mathbf{S}} I(X_m;X_j\,|\,C) - \frac{1}{|\mathbf{S}|}\sum_{X_j\in\mathbf{S}}\sum_{X_i\in\mathbf{S}; i\neq j} I(X_m;X_i\,|\,X_j) \right\} \tag{28}$$

This can be regarded as the JMI criterion in (3) ($\beta = \gamma = 1/|\mathbf{S}|$) with an additional consideration about the second-order interactions between the feature under consideration and the selected feature set. However, the problem in the above normalization is that the sum of second-order feature interaction terms $I(X_m;X_i\,|\,X_j)$'s is still so high that it may outweigh the importance of other terms. As a result, we propose to further normalize this term as

Form − 2:

$$\max_{X_m\in\mathbf{X}\setminus\mathbf{S}} \left\{ I(X_m;C) - \frac{1}{|\mathbf{S}|}\sum_{X_j\in\mathbf{S}} I(X_m;X_j) + \frac{1}{|\mathbf{S}|}\sum_{X_j\in\mathbf{S}} I(X_m;X_j\,|\,C) - \frac{1}{|\mathbf{S}||\mathbf{S}-1|}\sum_{X_j\in\mathbf{S}}\sum_{X_i\in\mathbf{S}; i\neq j} I(X_m;X_i\,|\,X_j) \right\} \tag{29}$$

Table 1
Data set description. The error rate is obtained by a linear SVM using all features.

Data	#Features (M)	#Instances (N)	#Classes	Instances–features ratio (N/M)	Problem scale ($M*N$)	Err(%)	Source
Wine	13	178	3	13.69	2314	3.04	[14]
Parkinsons	22	195	2	8.86	4290	12.93	[14]
Ionosphere	33	351	2	10.64	11,583	12.36	[14]
Breast	30	569	2	18.97	17,070	3.13	[14]
Lung	325	73	7	0.22	23,725	12.33	[5]
Segment	19	2310	7	121.58	43,890	6.36	[14]
Cardio	21	2126	3	101.24	44,646	10.73	[14]
Steel	27	1941	7	71.89	52,407	30.06	[14]
Musk	166	476	2	2.87	79,016	15.13	[14]
Waveform	21	5000	3	238.10	105,000	13.12	[14]
Arrhythmia	257	430	2	1.67	110,510	21.07	[14]
Colon	2000	62	2	0.03	124,000	17.74	[5]
Landsat	36	6435	6	178.75	231,660	13.60	[14]
Spambase	57	4601	2	80.72	262,257	9.72	[14]
Lymphoma	4026	96	9	0.02	386,496	3.12	[5]
Semeion	256	1593	10	6.22	407,808	6.26	[14]
Leukemia	7129	73	2	0.01	520,417	1.37	[5]
NCI60	9996	60	10	0.01	599,760	43.33	[5]

This normalization essentially aims to bring all the MI terms to the same scale. Note that the above criteria take into account the second-order feature interaction terms $I(X_m;X_i|X_j)$ which has never been explored in previous research to our knowledge.

4.1. Complexity analysis

We provide a complexity analysis for the newly designed criteria. Suppose the number of records in the data set is N, the number of features is M. Both mutual information $I(X;Y)$ and conditional mutual information $I(X;Y|Z)$ admits a time complexity of $O(N)$ since all the data points need to be visited for probability estimation.

Complexity of MIFS/MRMR/JMI and similar existing criteria: Suppose the number of features to be selected is k, then the complexity of MI-based feature selection algorithms, such as MRMR, MIFS, JMI and similar existing criteria are $O(k^2MN)$. Note that for improved efficiency, the redundancy and class-conditional redundancy terms could be cached in $M \times M$ tables for re-use.

Complexity of RelaxMRMR: Compared to MRMR, the time for RelaxMRMR is augmented by the time required for computing the second-order feature interaction terms. The time complexity for RelaxMRMR is $O(k^3MN)$. Again for improved efficiency, the second-order feature interaction terms can be cached in a $M \times M \times M$ table for re-use. RelaxMRMR is a generally more computationally intensive since more information is taken into account.

5. Experimental evaluation

In order to evaluate the performance of the newly proposed RelaxMRMR method, we performed an extensive experimental evaluation on a large number of real data sets detailed in Table 1. These data sets possess a wide range of characteristics, including varying numbers of features, instances and classes. The selected data sets represent a significant proportion of real world problems. For continuous numeric features, a discretization procedure is performed to categorize the original values into five equal-size bins. The implementation of RelaxMRMR in Matlab/C++ will be made available on our website, where the implementations for some most popular MI-based feature selection approaches are also available.[2]

Table 2
SVM error rate (%) comparison among different normalization forms, with Form-0 serving as the baseline.

Data	Form-0	Form-1	Form-2
Wine	9.10 ± 0.18	$7.23 \pm 0.21(-)$	$6.36 \pm 0.25(-)$
Parkinsons	16.86 ± 0.08	$15.77 \pm 0.17(-)$	$15.45 \pm 0.19(-)$
Ionosphere	15.65 ± 0.06	$13.19 \pm 0.02(-)$	$12.77 \pm 0.02(-)$
Breast	4.47 ± 0.02	$3.97 \pm 0.01(-)$	$3.73 \pm 0.01(-)$
Lung	28.85 ± 0.62	$22.58 \pm 0.64(-)$	$12.79 \pm 1.24(-)$
Segment	12.10 ± 0.94	$12.14 \pm 0.95(=)$	$10.67 \pm 1.01(-)$
Cardio	14.74 ± 0.12	$14.29 \pm 0.15(=)$	$13.30 \pm 0.12(-)$
Steel	38.93 ± 0.68	$37.39 \pm 0.74(-)$	$37.12 \pm 0.61(-)$
Musk	33.56 ± 0.06	$25.98 \pm 0.31(-)$	$25.50 \pm 0.32(-)$
Waveform	18.41 ± 0.52	$21.74 \pm 0.51(+)$	$18.03 \pm 0.55(-)$
Arrhythmia	24.73 ± 0.03	$25.65 \pm 0.02(+)$	$22.51 \pm 0.04(-)$
Colon	27.29 ± 0.65	$17.19 \pm 0.36(-)$	$12.61 \pm 0.44(-)$
Landsat	15.62 ± 0.25	$16.44 \pm 0.24(+)$	$15.95 \pm 0.25(+)$
Spambase	17.19 ± 0.22	$19.55 \pm 0.26(+)$	$13.97 \pm 0.30(-)$
Lymphoma	31.33 ± 0.11	$11.10 \pm 0.64(-)$	$9.04 \pm 0.63(-)$
Semeion	33.33 ± 1.17	$23.26 \pm 2.15(-)$	$29.93 \pm 1.56(-)$
Leukemia	9.53 ± 0.01	$5.56 \pm 0.03(-)$	$3.62 \pm 0.05(-)$
NCI60	85.77 ± 0.19	$69.70 \pm 0.15(-)$	$44.30 \pm 1.89(-)$
Win/Tie/Loss (for Form-0 vs. the alternative)	–	4/2/12	1/0/17

'+'/'−'/'=' indicates that Form-0 performs 'better'/'worse'/'equally well' compared to the competitor according to the t-test.

First, we evaluate the effectiveness of different normalization strategies and identify the best normalization approach. Based on this evaluation, we then compare the best-performing RelaxMRMR variant with other incremental MI-based methods in terms of effectiveness and efficiency. In addition, we also compare RelaxMRMR with some other representative non-incremental MI-based approaches and non MI-based approaches.

Our experimental protocol is as follows: for data sets with more than 50 features, we selected the top 50 features, while for lower dimensional data sets, all features are incrementally selected. Similarly to some previous research [12,13,9], for each feature set size, we employed a linear support vector machine (with the regularization parameter set to 1) to obtain the 10-fold cross-validation error rate (or leave-one-out validation error if the data

[2] http://vinhnguyenx.net/software.

Table 3
Error rate (%) comparison between RelaxMRMR and other incremental MI-based criteria.

Dataset	RelaxMRMR	MIM	MIFS(0.5)	MIFS(1)	MRMR	CIFE	JMI
SVM							
Wine	6.4 ± 0.3	5.9 ± 0.2(=)	6.6 ± 0.2(=)	8.6 ± 0.2(+)	6.4 ± 0.3(=)	9.1 ± 0.2(+)	6.2 ± 0.3(=)
Parkinsons	15.4 ± 0.2	15.2 ± 0.1(=)	14.1 ± 0.1(−)	16.2 ± 0.2(+)	15.1 ± 0.2(−)	15.2 ± 0.1(=)	14.8 ± 0.1(=)
Ionosphere	12.8 ± 0.0	17.2 ± 0.0(+)	13.4 ± 0.0(+)	13.3 ± 0.0(+)	13.4 ± 0.0(+)	16.5 ± 0.0(+)	16.7 ± 0.0(+)
Breast	3.7 ± 0.0	4.9 ± 0.0(+)	4.2 ± 0.0(+)	3.9 ± 0.0(+)	3.9 ± 0.0(+)	4.3 ± 0.0(+)	3.9 ± 0.0(=)
Lung	12.8 ± 1.2	19.8 ± 1.8(+)	12.3 ± 1.1(=)	14.9 ± 0.9(+)	12.9 ± 1.0(=)	26.8 ± 0.7(+)	13.5 ± 0.9(=)
Segment	10.7 ± 1.0	16.7 ± 1.6(+)	11.5 ± 1.0(+)	12.1 ± 1.0(+)	10.7 ± 1.0(=)	11.2 ± 1.0(+)	11.3 ± 1.0(+)
Cardio	13.3 ± 0.1	13.3 ± 0.1(=)	14.2 ± 0.1(+)	14.6 ± 0.1(+)	13.6 ± 0.1(=)	15.2 ± 0.1(+)	13.3 ± 0.1(=)
Steel	37.1 ± 0.6	41.2 ± 0.6(+)	37.8 ± 0.7(=)	37.9 ± 0.8(=)	38.2 ± 0.6(+)	39.0 ± 0.7(+)	40.4 ± 0.7(+)
Musk	25.5 ± 0.3	26.4 ± 0.2(+)	25.5 ± 0.4(=)	24.9 ± 0.4(=)	25.2 ± 0.3(=)	30.6 ± 0.1(+)	25.6 ± 0.2(=)
Waveform	18.0 ± 0.5	20.6 ± 0.8(+)	20.6 ± 0.5(+)	22.7 ± 0.6(+)	18.0 ± 0.5(=)	19.8 ± 0.4(+)	18.1 ± 0.5(=)
Arrhythmia	22.5 ± 0.0	23.4 ± 0.1(+)	24.1 ± 0.0(+)	24.7 ± 0.0(+)	23.2 ± 0.0(+)	25.7 ± 0.0(+)	23.0 ± 0.1(=)
Colon	12.6 ± 0.4	13.5 ± 0.2(+)	16.3 ± 0.4(+)	21.0 ± 0.2(+)	13.6 ± 0.4(+)	31.4 ± 0.3(+)	14.6 ± 0.6(+)
Landsat	16.0 ± 0.3	16.0 ± 0.2(=)	16.5 ± 0.2(+)	16.6 ± 0.3(+)	15.9 ± 0.3(=)	15.5 ± 0.3(−)	15.6 ± 0.2(−)
Spambase	14.0 ± 0.3	13.8 ± 0.3(=)	17.9 ± 0.3(+)	20.2 ± 0.4(+)	13.9 ± 0.3(=)	20.2 ± 0.2(+)	13.9 ± 0.3(=)
Lymphoma	9.0 ± 0.6	16.3 ± 0.8(+)	7.9 ± 0.6(−)	12.2 ± 0.4(+)	8.4 ± 0.6(−)	29.2 ± 0.2(+)	8.8 ± 0.5(=)
Semeion	29.9 ± 1.6	39.1 ± 2.5(+)	22.1 ± 2.4(−)	23.4 ± 2.2(−)	32.6 ± 1.6(+)	35.4 ± 1.2(+)	34.3 ± 1.6(+)
Leukemia	3.6 ± 0.1	4.5 ± 0.0(+)	10.6 ± 0.1(+)	9.6 ± 0.0(+)	3.6 ± 0.1(=)	12.5 ± 0.1(+)	3.8 ± 0.0(=)
NCI60	44.3 ± 1.9	50.5 ± 1.5(+)	51.2 ± 1.3(+)	60.3 ± 0.7(+)	45.3 ± 2.0(=)	86.5 ± 0.2(+)	45.6 ± 1.9(+)
Win/Tie/Loss	–	13/5/0	11/4/3	15/2/1	6/10/2	16/1/1	6/11/1
Naive Bayes							
Wine	14.8 ± 2.1	17.3 ± 2.3(+)	15.0 ± 2.2(=)	15.0 ± 2.1(=)	15.2 ± 2.0(=)	15.9 ± 2.1(=)	14.9 ± 2.1(=)
Parkinsons	19.0 ± 0.4	19.8 ± 0.2(=)	20.6 ± 0.3(+)	20.0 ± 0.3(+)	18.7 ± 0.3(−)	21.0 ± 0.3(+)	19.3 ± 0.2(=)
Ionosphere	27.5 ± 0.2	27.8 ± 0.3(=)	28.6 ± 0.1(+)	27.7 ± 0.1(=)	29.4 ± 0.1(+)	31.4 ± 0.1(+)	29.3 ± 0.2(+)
Breast	26.3 ± 0.2	33.6 ± 0.2(+)	23.6 ± 0.3(−)	23.5 ± 0.2(−)	27.1 ± 0.3(=)	24.5 ± 0.2(−)	31.3 ± 0.1(+)
Lung	16.0 ± 2.7	27.5 ± 3.2(+)	15.5 ± 1.8(=)	15.4 ± 1.6(=)	15.5 ± 1.8(=)	32.5 ± 1.1(+)	16.9 ± 2.3(=)
Segment	27.5 ± 3.0	37.3 ± 4.7(+)	27.5 ± 3.0(=)	30.6 ± 2.8(+)	27.5 ± 3.0(=)	31.0 ± 2.8(+)	28.8 ± 3.0(=)
Cardio	17.0 ± 0.1	18.5 ± 0.0(+)	17.2 ± 0.1(=)	18.1 ± 0.0(+)	17.1 ± 0.1(=)	18.9 ± 0.0(+)	18.2 ± 0.0(+)
Steel	44.5 ± 1.2	47.4 ± 0.4(+)	41.7 ± 0.5(−)	44.7 ± 1.0(=)	44.4 ± 1.1(=)	46.1 ± 0.9(+)	45.8 ± 0.6(=)
Musk	28.7 ± 0.2	31.7 ± 0.1(+)	32.4 ± 0.2(+)	30.6 ± 0.3(+)	29.2 ± 0.3(+)	34.1 ± 0.0(+)	30.5 ± 0.1(+)
Waveform	23.7 ± 1.3	27.4 ± 1.4(+)	24.5 ± 1.1(=)	27.4 ± 1.3(+)	22.9 ± 1.2(=)	24.0 ± 1.1(=)	23.3 ± 1.2(=)
Arrhythmia	24.0 ± 0.1	28.7 ± 0.1(+)	25.1 ± 0.1(+)	28.1 ± 0.1(+)	23.8 ± 0.1(=)	34.2 ± 0.0(+)	29.0 ± 0.1(+)
Colon	10.8 ± 0.3	11.9 ± 0.1(+)	16.4 ± 0.2(+)	14.4 ± 0.2(+)	12.3 ± 0.3(+)	24.3 ± 0.1(+)	12.7 ± 0.4(+)
Landsat	27.1 ± 0.8	30.4 ± 1.4(+)	27.7 ± 0.8(+)	28.4 ± 0.8(+)	27.4 ± 0.9(+)	27.5 ± 0.8(+)	27.0 ± 0.8(=)
Spambase	10.6 ± 0.6	14.0 ± 0.7(+)	10.6 ± 0.5(=)	12.0 ± 0.3(+)	11.1 ± 0.5(+)	12.8 ± 0.3(+)	12.4 ± 0.6(+)
Lymphoma	11.5 ± 1.9	20.5 ± 1.3(+)	12.7 ± 1.0(+)	19.6 ± 1.1(+)	10.5 ± 1.6(−)	42.0 ± 0.1(+)	10.9 ± 1.3(=)
Semeion	38.0 ± 1.5	48.5 ± 2.8(+)	28.8 ± 2.7(−)	29.7 ± 2.3(−)	40.9 ± 1.7(+)	47.7 ± 0.9(+)	43.6 ± 1.9(+)
Leukemia	3.2 ± 0.9	4.7 ± 1.0(+)	4.3 ± 0.4(=)	8.9 ± 0.2(+)	1.5 ± 0.2(=)	16.2 ± 0.2(+)	2.2 ± 0.7(=)
NCI60	40.2 ± 2.7	43.1 ± 2.4(+)	41.5 ± 2.0(=)	54.4 ± 0.9(+)	41.4 ± 3.0(+)	85.7 ± 0.1(+)	37.7 ± 2.6(−)
Win/Tie/Loss	–	16/2/0	7/8/3	12/4/2	7/9/2	15/2/1	8/9/1
KNN							
Wine	5.9 ± 0.2	5.8 ± 0.3(=)	6.2 ± 0.3(=)	7.7 ± 0.2(+)	5.8 ± 0.2(=)	8.8 ± 0.2(+)	5.8 ± 0.2(=)
Parkinsons	8.6 ± 0.1	10.2 ± 0.1(+)	9.0 ± 0.1(=)	9.2 ± 0.1(+)	8.8 ± 0.1(=)	9.0 ± 0.1(+)	9.4 ± 0.1(+)
Ionosphere	13.1 ± 0.1	14.1 ± 0.0(+)	12.8 ± 0.0(=)	12.7 ± 0.1(=)	12.8 ± 0.0(=)	14.4 ± 0.1(+)	13.2 ± 0.0(=)
Breast	3.5 ± 0.0	5.0 ± 0.0(+)	4.6 ± 0.0(+)	4.3 ± 0.0(+)	3.6 ± 0.0(=)	4.9 ± 0.0(+)	4.2 ± 0.0(+)
Lung	13.1 ± 0.9	26.8 ± 1.0(+)	14.0 ± 0.9(+)	20.1 ± 0.8(+)	14.4 ± 0.9(+)	34.8 ± 0.2(+)	15.7 ± 0.8(+)
Segment	5.8 ± 0.7	9.0 ± 0.8(+)	6.0 ± 0.7(+)	6.5 ± 0.7(+)	5.8 ± 0.7(=)	5.9 ± 0.7(=)	5.8 ± 0.7(=)
Cardio	10.1 ± 0.2	9.7 ± 0.2(−)	13.0 ± 0.3(+)	13.6 ± 0.3(+)	11.0 ± 0.3(+)	11.0 ± 0.1(+)	9.4 ± 0.2(−)
Steel	33.4 ± 1.3	36.1 ± 1.2(+)	34.7 ± 1.3(+)	34.4 ± 1.3(+)	34.2 ± 1.3(+)	34.0 ± 1.2(=)	34.0 ± 1.2(=)
Musk	21.1 ± 0.4	25.0 ± 0.2(+)	22.0 ± 0.4(+)	22.2 ± 0.5(+)	21.2 ± 0.4(=)	18.3 ± 0.3(−)	23.0 ± 0.2(+)
Waveform	23.4 ± 0.6	26.7 ± 1.0(+)	27.8 ± 0.4(+)	30.2 ± 0.6(+)	23.3 ± 0.6(=)	26.7 ± 0.4(+)	23.4 ± 0.6(=)
Arrhythmia	25.3 ± 0.1	28.5 ± 0.1(+)	25.7 ± 0.1(=)	28.0 ± 0.1(+)	26.4 ± 0.1(+)	32.8 ± 0.1(+)	28.2 ± 0.0(+)
Colon	14.7 ± 0.0	15.7 ± 0.1(+)	24.2 ± 0.1(+)	20.7 ± 0.1(+)	14.6 ± 0.1(=)	26.7 ± 0.1(+)	16.6 ± 0.1(+)
Landsat	12.0 ± 0.7	12.9 ± 0.7(+)	12.4 ± 0.7(+)	12.6 ± 0.7(+)	12.0 ± 0.7(=)	12.2 ± 0.7(+)	12.3 ± 0.7(+)
Spambase	14.4 ± 1.3	14.3 ± 1.4(=)	18.1 ± 1.1(+)	20.4 ± 1.1(+)	14.3 ± 1.4(=)	15.1 ± 0.9(=)	14.4 ± 1.4(=)
Lymphoma	12.4 ± 0.5	18.0 ± 0.6(+)	13.4 ± 0.4(+)	20.5 ± 0.2(+)	9.8 ± 0.7(−)	40.3 ± 0.4(+)	11.8 ± 0.5(−)
Semeion	34.3 ± 2.0	43.9 ± 3.3(+)	26.0 ± 3.1(−)	28.8 ± 2.6(−)	37.4 ± 2.2(+)	37.2 ± 1.6(+)	39.2 ± 2.3(+)
Leukemia	2.4 ± 0.0	4.8 ± 0.0(+)	18.1 ± 0.3(+)	19.5 ± 0.4(+)	3.0 ± 0.0(+)	14.6 ± 0.3(+)	2.7 ± 0.0(+)
NCI60	42.2 ± 1.4	51.0 ± 0.9(+)	49.8 ± 1.2(+)	55.9 ± 1.4(+)	46.6 ± 1.1(+)	84.6 ± 0.1(+)	49.1 ± 1.0(+)
Win/Tie/Loss	-	15/2/1	13/4/1	16/1/1	7/10/1	14/3/1	10/6/2

'+'/'−'/'=' indicates that RelaxMRMR performs 'better'/'worse'/'equally well' compared to the competitor according to the t-test.

set contains less than 100 instances). Additionally, the same statistics are also collected from two other classifiers, namely Naive Bayes (NB) and kNN classifier (k=3). As such, given a data set, a certain classifier and a specific feature selection method, a plot of the cross-validation error rate vs. the number of features can be drawn and we can also compute the *mean* ± *standard deviation* %-error rate across a range of feature set size (from 1 to the maximum number of selected features).

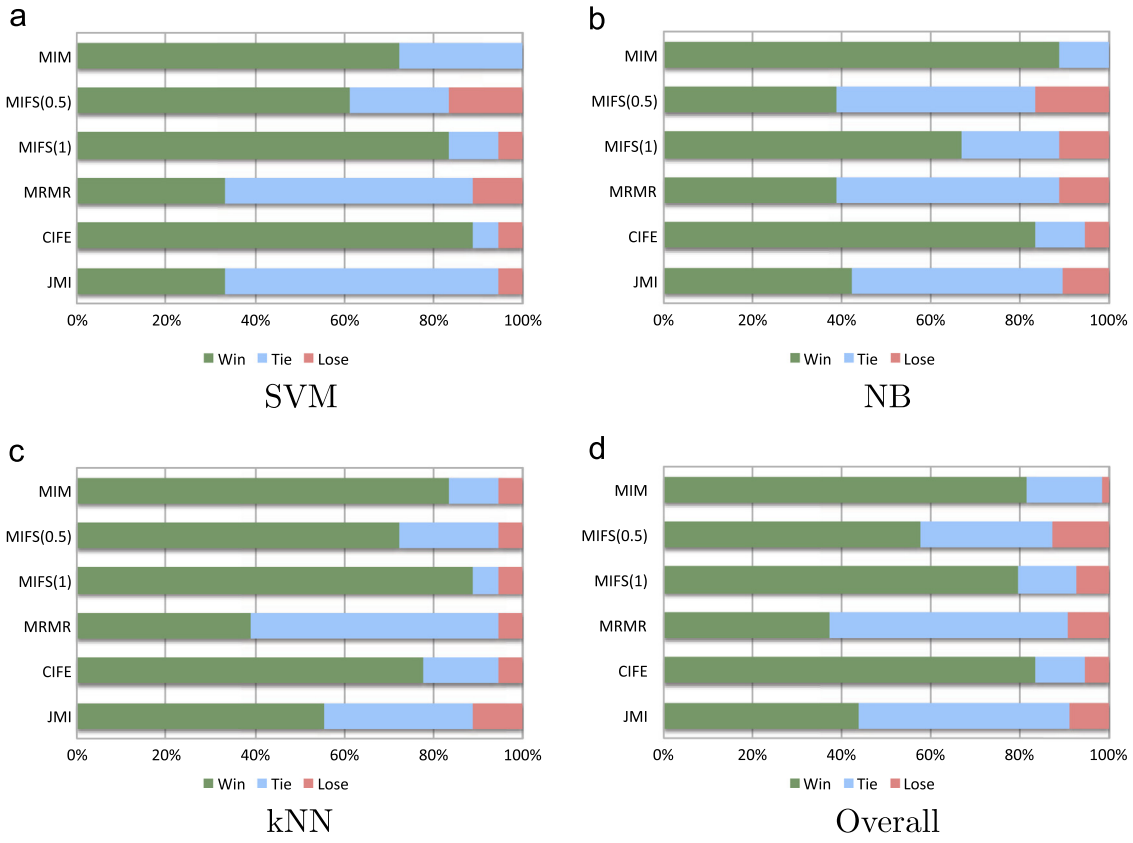

Fig. 2. Performance comparison of RelaxMRMR to other incremental MI-based criteria. Win/Tie/Loss means RelaxMRMR performs 'better'/'equally-well'/'worse' than the alternatives. (a) SVM, (b) NB, (c) kNN, (d) Overall.

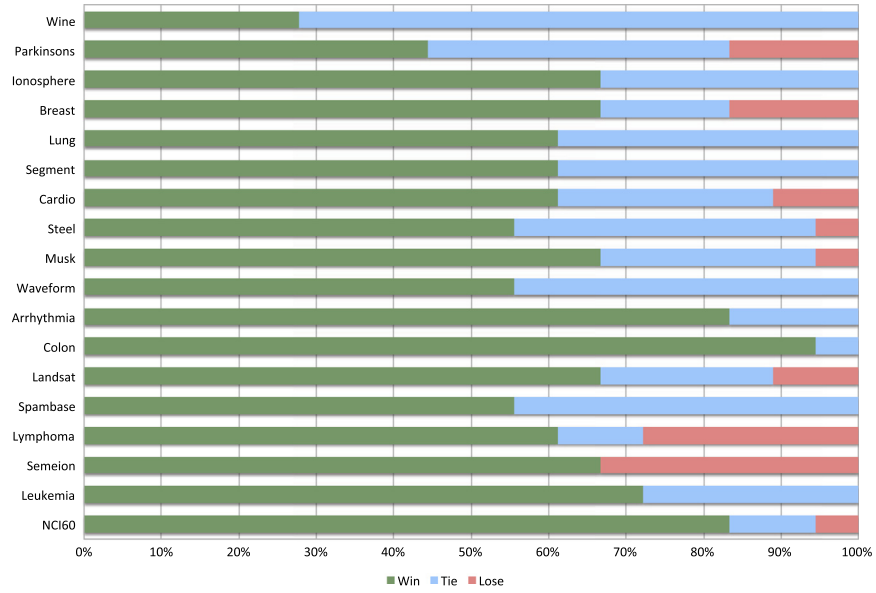

Fig. 3. Detailed performance comparison of RelaxMRMR to other incremental MI-based criteria on each data set (sorted by *#Features* × *#Instances*) across all classifiers. Win/Tie/Loss means RelaxMRMR performs 'better'/'equally well'/'worse' than other methods.

5.1. Normalization

We tested Form-1, Form-2 and the un-normalized form (Form-0) of RelaxMRMR on all data sets. To determine which normalization form performs better overall, following Herman et al. [13], the one-sided paired t-test at 5% significance level was used to compare Form-1 and Form-2 with the baseline Form-0. The experiment results of SVM are shown in Table 2 where we use

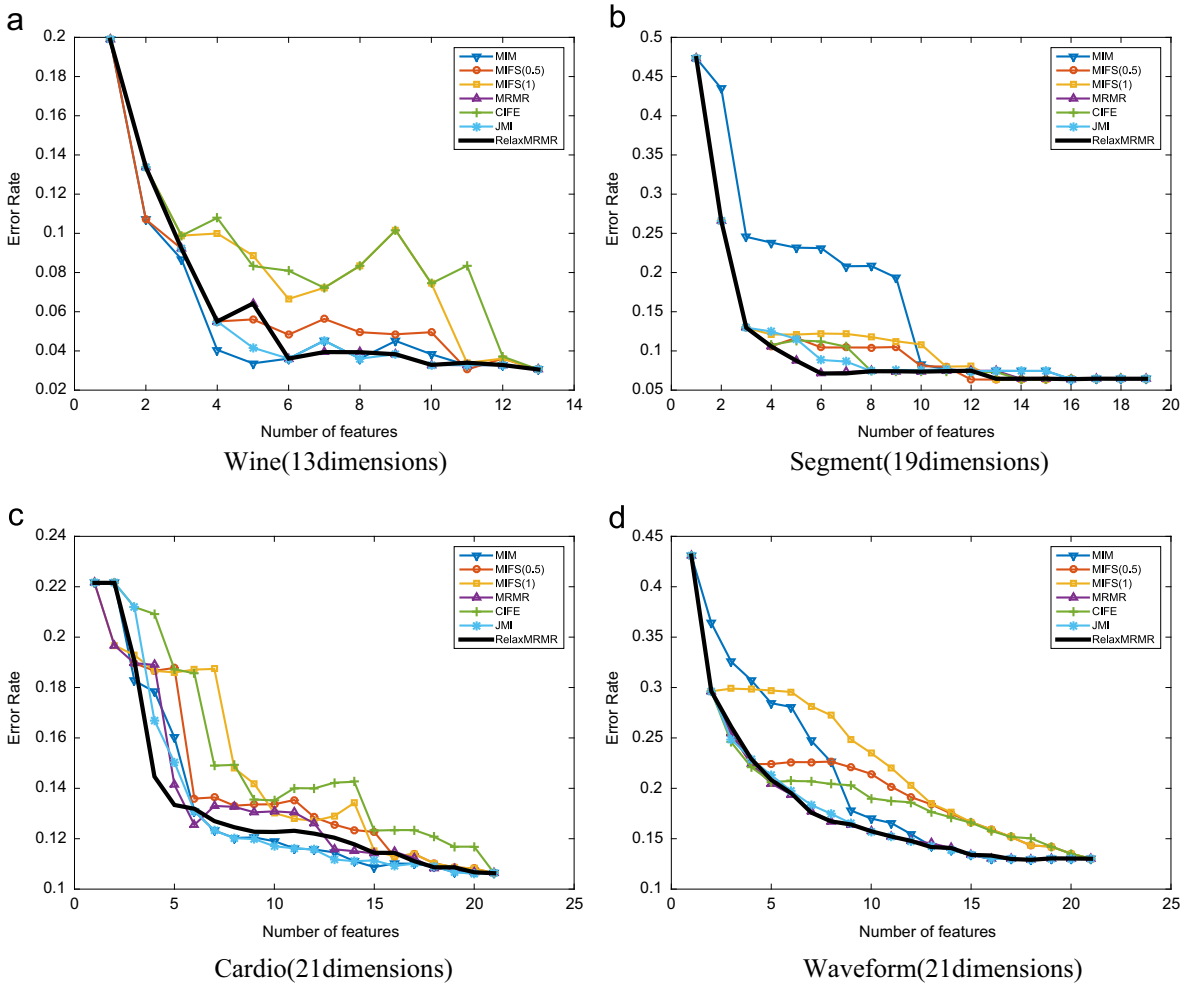

Fig. 4. Performance comparison on low-dimensional data sets with SVM (best viewed in color).

'+'/'−'/'=' to indicate that Form-0 performs 'better'/'worse'/'equally well' compared to the two other forms. Although not reported, we observed similar results with NB and kNN.

As can be seen from this table, normalization does improve the performance of RelaxMRMR. In most cases, the unnormalized form is outperformed by both Form-1 and Form-2. In addition, using the same testing procedure, the win/tie/loss counts of Form-2 vs. Form-1 is 16/1/1. Thus clearly, Form-2 consistently performs better than Form-1. This experimental result verifies the effectiveness of normalizing different MI quantities to a similar scale. This normalization prevents the algorithm from being largely biased towards a particular factor. Our finding is in concordance with previous research. For example, Ref. [3] showed that MRMR usually outperforms MIFS while JMI often outperforms CIFE. Both the two winning methods, MRMR and JMI, follow the same normalization strategy that brings every term in the objective into a similar scale, while MIFS and CIFE employ unnormalized objectives.

5.2. Comparison with incremental MI-based approaches

We compared the normalized RelaxMRMR (Form-2) with other existing MI-based approaches that select features in an *incremental* fashion, including Mutual Information Maximisation (MIM), also known as the Maximum Relevance criterion ($\beta = 0$ in (2)) [5], MIFS with $\beta = 0.5$ and $\beta = 1$ in (2) [4], MRMR [5], CIFE [8] and JMI [6]. The one-sided paired t-test was used to compare RelaxMRMR against other methods. We used '+'/'−'/'=' to indicate that RelaxMRMR performs 'better'/'worse'/'equally well' compared to the competitor. The result is shown in Table 3 and summarized in Fig. 2. Additionally, the performance with respect to the individual data sets is provided in Fig. 3.

5.2.1. Overall effectiveness

In general, compared with existing incremental MI-based methods, RelaxMRMR performs considerably well. Specifically, in more than 50% of the cases, the proposed approach yields better effectiveness than all other approaches, while there is only ~10% of the cases where one of the competitors wins. In the remaining one-third cases, the performances of RelaxMRMR and the other approaches are similar. Again, it should be noticed that algorithms without a good balancing between the relevancy and redundancy (e.g. MIFS with $\beta = 1$ and CIFE) usually provide worse performance than others. We observed similar results across all the three classifiers.

5.2.2. Effectiveness with respect to the number of features

As shown in Figs. 4 and 5, on low-dimensional data sets, there is no significant difference among different feature selection approaches. Since the number of features in a data set is limited to a small number, the relationship among these features is relatively simpler than that of high-dimensional data. As a result, restrictive

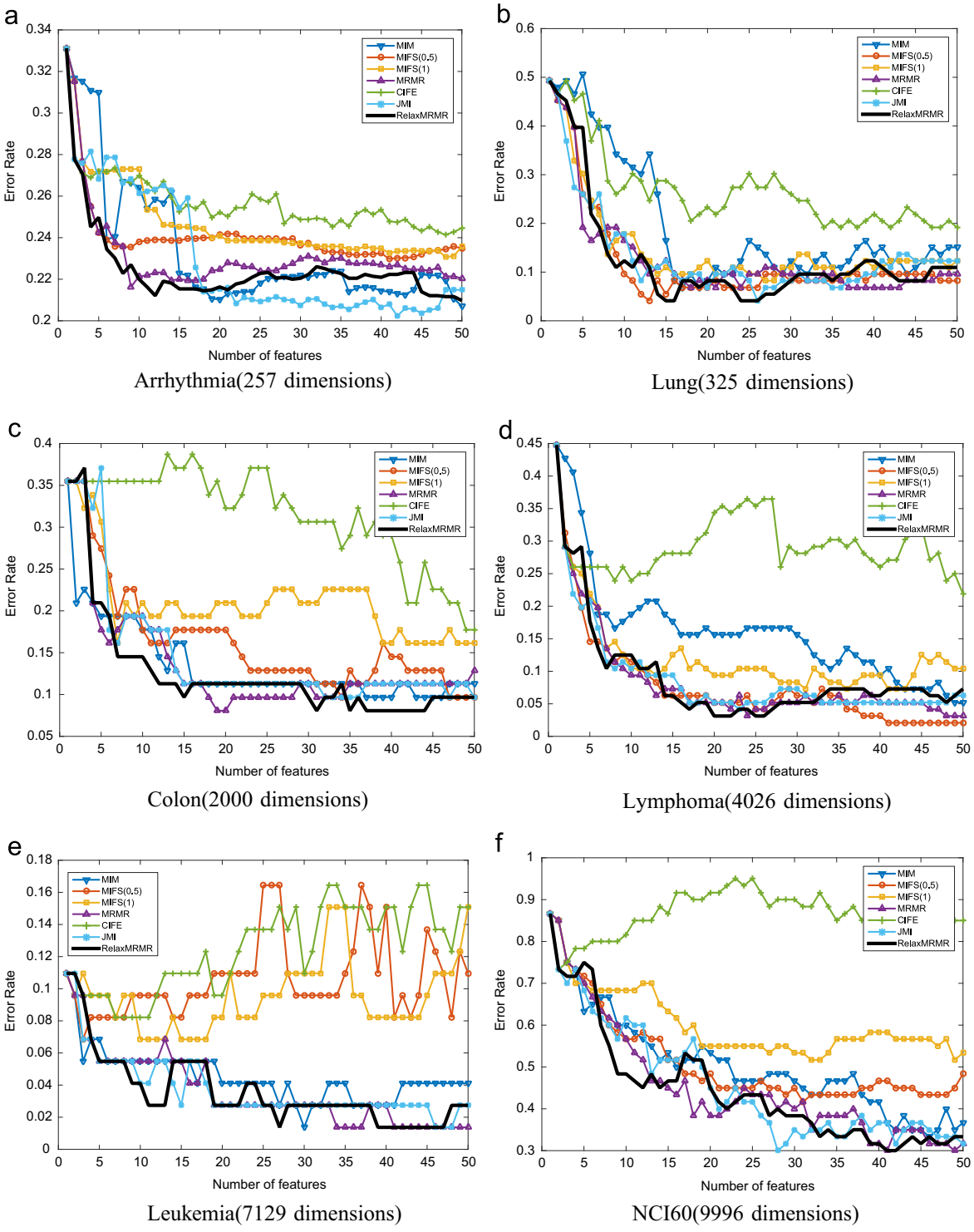

Fig. 5. Performance comparison on high-dimensional data sets with SVM (best viewed in color).

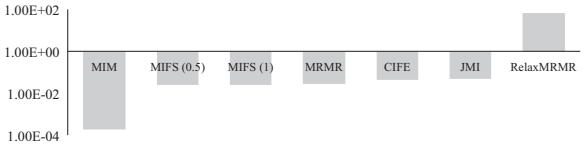

Fig. 6. Run time (seconds) comparison between RelaxMRMR and other MI-based algorithms, measured by the average time for selecting one feature from the Gisette data set.

assumptions can be applied and simple criteria such as MIM, MIFS and MRMR may already be sufficiently capable to achieve reasonable performance. As a matter of fact, for these low-dimensional data sets, it is more important to balance the relative importance between relevancy and redundancy, rather than introducing advanced terms such as class-relevant redundancy or high-order interactions between features.

On the other hand, in high-dimensional data, the underlying dependency structure between features within the data set is

dramatically more complex. In this situation, strong assumptions, such as pairwise independency, are unrealistic (for example, MIM tends to perform really badly on data sets with more than a hundred features). In contrast, the proposed RelaxMRMR approach, which posits the weakest assumptions among all MI-based algorithms, gains significant effectiveness due to the fact that it takes into account more underlying relationships among features in the data set.

5.2.3. Effectiveness with respect to the data set size

The effectiveness of RelaxMRMR is not only affected by the dimensionality of the feature set, but also the data size. Large data size is crucial for high-dimensional MI estimation. High-dimensional MI that is estimated based on a small amount of data is less reliable and may affect the performance of a high-dimensional MI-based feature selection method. However, even though the lack of data is a serious challenge, surprisingly it did not significantly offset the effectiveness of RelaxMRMR on high-dimensional data sets. In fact, for data sets with less than 100 data points but with a considerable number of features, i.e., Colon, Lymphoma, Leukemia and NCI60, in approximately 80% of the cases RelaxMRMR performs better than its competitors.

5.2.4. Efficiency

We tested the efficiency of RelaxMRMR compared to other incremental MI-based approaches. As the theoretical analysis in Section 4.1 has suggested, the complexity of RelaxMRMR is $O(k^3MN)$ compared to $O(k^2MN)$ of other MI-based approaches. RelaxMRMR is thus generally more computationally demanding. However to our observation, this difference is not practically significant for small to medium data sets. To gain a concrete idea of wall clock processing time, we tested the algorithms on a large high-dimensional data sets, namely Gisette from the NIPS feature selection competition [15] of 5000 features and 6000 data points. The result is shown in Fig. 6.

On this high-dimensional data set, RelaxMRMR is considerably more expensive. Nevertheless, we expect that the time complexity should not be a major deterrent to the practicality of RelaxMRMR. There are two arguments to support this claim. First, there are many applications where the data collecting time is far more than the time required for data mining tasks such as feature selection (e.g., days to months for data collection vs. hours for data mining). In these cases, it is justifiable to spend significant amounts of time for data processing and the improved performance brought about by RelaxMRMR will be worth the effort. Second, commodity multi-core systems are common nowadays, and it is straightforward to parallelize RelaxMRMR to harness this parallel processing power. Towards this end, we tested a parallel version of RelaxMRMR, where the high-order feature interactions terms are computed in parallel on a 16-core and a 32-core system. The effectiveness of parallelization can be clearly observed in Fig. 7.

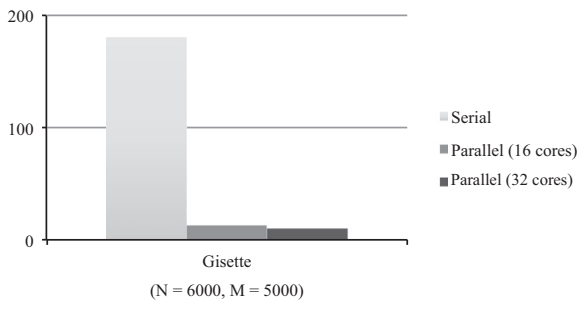

Fig. 7. Run time (seconds) comparison between serial and parallel RelaxMRMR, measured by the average time for selecting one feature from the Gisette data set.

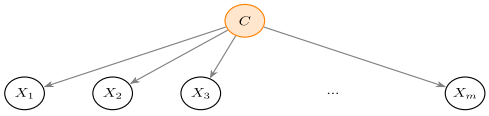

Fig. 8. Bayesian network representation of the Naive Bayes classifier. Each node (feature) has only 1 parent, which is C.

5.3. Comparison with other feature selection methods

In the previous section, we discussed the performance of RelaxMRMR compared to some well-known *incremental* MI-based feature selection methods. To provide a more comprehensive picture, we also compared RelaxMRMR against some non-incremental MI-based approaches and non-MI based approaches. In particular, we chose two global MI-based approaches that formulate feature selection as a global optimization problem, namely Quadratic Programming Feature Selection (QPFS) [16] and SpecCMI [9], and two representative non MI-based techniques, namely spectral feature selection [17] and ReliefF [18]. The result of these experiments is shown in Table 4. Overall, RelaxMRMR exhibits strong performance compared to other approaches. It is noted that while QPFS and SpecCMI use a global optimization approach, their objective functions are similar to ones employed by the incremental MRMR and JMI approaches respectively. None of these methods made use of the second order feature interaction terms.

6. Related work and discussion

In this section, we highlight the analogy between the MI-based feature selection problem and the related problem of building Naive Bayes classifier and its independence-assumptions-relaxed variants. Naive Bayes classifiers make a strong independence assumption, that all the features are conditionally independent given the value of the class C, which is in fact Assumption 3.

The dependency between features and the class variable can be represented intuitively by means of a Bayesian network, as in Fig. 8, wherein a node (feature) is probabilistically independent of all its non-descendants, given its parents. Despite its strong independence assumption, Naive Bayes classifiers often perform well in practice. Nevertheless, while it is known that some violation of the independency assumption do not matter, many others do affect the performance of Naive Bayes classifiers badly [19]. To this end, there have been a rich body of work on relaxing the strong independence assumption for Naive Bayes. Two of the popular approaches are the Tree-Augmented Naive Bayes (TAN) [20], and the Averaged One-Dependence Estimators (AODE) [19].

In TAN, the conditional independence assumption is relaxed, allowing each feature X_m to be independent of all other features, given C and at most another feature $p(X_m)$, called its parent, i.e., $P(X_m|X_1,\ldots,X_{m-1},C)=P(X_m|C,p(X_m))$. The conditional mutual information is used to select the parent. The Bayesian network structure of TAN is illustrated in Fig. 9.

In AODE, instead of learning the parent for each feature, the classifier is built by aggregating all 1-dependence classifiers. In each of these 1-dependence classifier, one feature is selected to be the parent for all other features. Each feature in turn plays the role of the parent. The structure of these 1-dependence classifiers are

Table 4
Error rate (%) comparison between RelaxMRMR and QPFS, SpecCMI, Spectral and ReliefF.

Dataset	RelaxMRMR	QPFS	SpecCMI	Spectral	ReliefF
SVM					
Wine	6.4 ± 0.3	5.8 ± 0.2(−)	7.9 ± 0.8(=)	6.0 ± 0.2(=)	12.9 ± 1.7(+)
Parkinsons	15.4 ± 0.2	13.6 ± 0.1(−)	14.8 ± 0.2(−)	15.3 ± 0.1(=)	14.7 ± 0.2(−)
Ionosphere	12.8 ± 0.0	15.6 ± 0.1(+)	17.7 ± 0.2(+)	14.1 ± 0.0(+)	18.5 ± 0.8(+)
Breast	3.7 ± 0.0	3.9 ± 0.0(=)	4.3 ± 0.0(+)	4.6 ± 0.0(+)	5.6 ± 0.5(+)
Lung	12.8 ± 1.2	11.9 ± 1.8(=)	18.1 ± 1.5(+)	19.3 ± 1.6(+)	22.8 ± 1.4(+)
Segment	10.7 ± 1.0	10.8 ± 1.0(=)	11.1 ± 1.1(+)	19.2 ± 3.9(+)	15.0 ± 2.0(+)
Cardio	13.3 ± 0.1	12.7 ± 0.1(−)	13.3 ± 0.1(=)	12.4 ± 0.1(−)	13.2 ± 0.1(=)
Steel	37.1 ± 0.6	37.3 ± 0.5(=)	40.1 ± 0.8(+)	39.7 ± 0.5(+)	40.8 ± 0.8(+)
Musk	25.5 ± 0.3	25.0 ± 0.3(−)	23.4 ± 0.3(−)	22.0 ± 0.1(−)	28.7 ± 0.6(+)
Waveform	18.0 ± 0.5	19.2 ± 0.9(+)	19.1 ± 0.9(+)	20.7 ± 0.9(+)	19.3 ± 0.9(+)
Arrhythmia	22.5 ± 0.0	24.7 ± 0.1(+)	24.2 ± 0.1(+)	23.3 ± 0.1(+)	24.0 ± 0.3(+)
Colon	12.6 ± 0.4	13.2 ± 0.2(=)	12.6 ± 0.1(=)	13.5 ± 0.1(=)	17.5 ± 0.1(+)
Landsat	16.0 ± 0.3	16.5 ± 0.5(=)	21.6 ± 1.4(+)	16.0 ± 0.2(=)	21.0 ± 1.3(+)
Spambase	14.0 ± 0.3	14.1 ± 0.2(=)	13.8 ± 0.3(−)	12.8 ± 0.1(−)	15.0 ± 0.3(+)
Lymphoma	9.0 ± 0.6	10.1 ± 0.7(+)	24.1 ± 0.9(+)	14.6 ± 0.5(+)	13.1 ± 1.0(+)
Semeion	29.9 ± 1.6	29.4 ± 3.1(=)	38.1 ± 2.4(+)	40.0 ± 2.4(+)	46.5 ± 2.7(+)
Leukemia	3.6 ± 0.1	4.4 ± 0.0(+)	5.0 ± 0.1(+)	5.0 ± 0.0(+)	4.7 ± 0.0(+)
NCI60	44.3 ± 1.9	N/A	52.8 ± 2.8(+)	53.1 ± 5.0(+)	56.7 ± 1.8(+)
Win/Tie/Loss	–	5/8/4	12/3/3	11/4/3	16/1/1
Naive Bayes					
Wine	14.8 ± 2.1	16.1 ± 2.1(+)	21.1 ± 4.1(+)	17.1 ± 2.3(+)	21.0 ± 3.7(+)
Parkinsons	19.0 ± 0.4	19.0 ± 0.2(=)	21.8 ± 0.2(+)	22.5 ± 0.1(+)	21.0 ± 0.2(+)
Ionosphere	27.5 ± 0.2	29.7 ± 0.1(+)	31.1 ± 0.2(+)	29.8 ± 0.2(+)	29.5 ± 0.1(+)
Breast	26.3 ± 0.2	30.1 ± 0.3(+)	27.8 ± 0.3(+)	33.4 ± 0.2(+)	28.4 ± 0.3(+)
Lung	16.0 ± 2.7	14.0 ± 2.4(−)	22.4 ± 1.8(+)	25.8 ± 3.4(+)	23.5 ± 2.8(+)
Segment	27.5 ± 3.0	27.4 ± 3.5(=)	28.7 ± 3.4(=)	34.4 ± 2.9(+)	31.5 ± 3.0(=)
Cardio	17.0 ± 0.1	17.4 ± 0.0(=)	18.6 ± 0.0(+)	18.3 ± 0.0(+)	20.3 ± 0.0(+)
Steel	44.5 ± 1.2	44.8 ± 0.9(=)	45.1 ± 0.6(=)	46.9 ± 0.5(+)	45.4 ± 0.6(=)
Musk	28.7 ± 0.2	30.7 ± 0.1(+)	30.5 ± 0.2(+)	29.6 ± 0.3(+)	33.8 ± 0.3(+)
Waveform	23.7 ± 1.3	26.3 ± 2.1(+)	26.1 ± 2.2(+)	27.2 ± 1.4(+)	26.2 ± 2.1(+)
Arrhythmia	24.0 ± 0.1	26.7 ± 0.3(+)	30.5 ± 0.1(+)	27.6 ± 0.1(+)	26.6 ± 0.2(+)
Colon	10.8 ± 0.3	11.9 ± 0.1(+)	14.3 ± 0.3(+)	16.3 ± 0.8(+)	35.4 ± 0.3(+)
Landsat	27.1 ± 0.8	29.2 ± 1.5(+)	38.2 ± 4.9(+)	31.1 ± 1.4(+)	40.1 ± 4.2(+)
Spambase	10.6 ± 0.6	9.8 ± 0.4(−)	12.2 ± 0.6(+)	12.9 ± 0.6(+)	18.0 ± 0.6(+)
Lymphoma	11.5 ± 1.9	13.5 ± 1.4(+)	23.5 ± 1.0(+)	18.7 ± 0.8(+)	18.2 ± 1.6(+)
Semeion	38.0 ± 1.5	33.5 ± 2.6(−)	47.5 ± 2.7(+)	49.9 ± 2.8(+)	59.7 ± 1.8(+)
Leukemia	3.2 ± 0.9	3.9 ± 0.8(+)	3.3 ± 0.9(=)	16.0 ± 2.4(+)	20.5 ± 1.4(+)
NCI60	40.2 ± 2.7	N/A	46.4 ± 2.6(+)	50.2 ± 5.6(+)	64.8 ± 2.1(+)
Win/Tie/Loss	–	10/4/3	15/3/0	18/0/0	16/2/0
KNN					
Wine	5.9 ± 0.2	5.8 ± 0.3(=)	7.6 ± 0.8(=)	6.0 ± 0.2(=)	9.8 ± 1.0(+)
Parkinsons	8.6 ± 0.1	9.8 ± 0.1(+)	9.5 ± 0.2(=)	9.7 ± 0.2(=)	9.9 ± 0.5(=)
Ionosphere	13.1 ± 0.1	13.9 ± 0.0(+)	14.0 ± 0.0(+)	14.0 ± 0.1(+)	14.3 ± 0.0(+)
Breast	3.5 ± 0.0	4.3 ± 0.0(+)	4.4 ± 0.0(+)	4.8 ± 0.0(+)	5.6 ± 0.6(+)
Lung	13.1 ± 0.9	13.0 ± 1.9(=)	24.1 ± 1.6(+)	34.2 ± 1.9(+)	25.8 ± 2.4(+)
Segment	5.8 ± 0.7	5.7 ± 0.7(=)	6.1 ± 0.7(=)	12.3 ± 3.9(+)	8.1 ± 1.4(+)
Cardio	10.1 ± 0.2	10.2 ± 0.2(=)	9.8 ± 0.2(−)	9.7 ± 0.2(=)	10.0 ± 0.1(=)
Steel	33.4 ± 1.3	33.7 ± 1.3(=)	36.0 ± 1.3(+)	39.6 ± 0.9(+)	35.6 ± 1.3(+)
Musk	21.1 ± 0.4	21.4 ± 0.4(=)	18.4 ± 0.5(−)	24.0 ± 0.2(+)	16.7 ± 0.4(−)
Waveform	23.4 ± 0.6	24.4 ± 0.9(=)	24.3 ± 0.9(=)	26.6 ± 0.9(+)	24.6 ± 0.9(+)
Arrhythmia	25.3 ± 0.1	29.5 ± 0.9(+)	28.9 ± 0.0(+)	28.6 ± 0.1(+)	27.5 ± 0.1(+)
Colon	14.7 ± 0.0	15.1 ± 0.1(=)	16.1 ± 0.0(+)	16.4 ± 0.1(+)	15.6 ± 0.1(+)
Landsat	12.0 ± 0.7	13.0 ± 1.0(=)	18.2 ± 2.1(+)	12.6 ± 0.5(+)	17.0 ± 1.7(+)
Spambase	14.4 ± 1.3	16.6 ± 1.7(+)	14.1 ± 1.4(=)	15.6 ± 1.7(+)	17.2 ± 1.3(+)
Lymphoma	12.4 ± 0.5	12.5 ± 0.7(=)	26.2 ± 0.7(+)	15.3 ± 0.4(+)	15.2 ± 1.1(+)
Semeion	34.3 ± 2.0	35.2 ± 3.6(=)	42.7 ± 3.2(+)	46.1 ± 3.6(+)	49.6 ± 3.1(+)
Leukemia	2.4 ± 0.0	3.5 ± 0.0(+)	5.0 ± 0.0(+)	5.3 ± 0.0(+)	6.8 ± 0.0(+)
NCI60	42.2 ± 1.4	N/A	55.7 ± 1.2(+)	53.6 ± 2.6(+)	52.6 ± 0.4(+)
Win/Tie/Loss	-	6/11/0	11/5/2	15/3/0	15/2/1

'+'/'−'/'=' indicates that RelaxMRMR performs 'better'/'worse'/'equally well' compared to the competitor according to the t-test.
N/A denotes QPFS returning a 'non-convexity' error.

illustrated in Fig. 10. In AODE, the joint distribution is factorized as:

$$P(\mathbf{X}, C) = \frac{1}{n}\sum_{i=1}^{n} P(X_i, C) \prod_{j=1, j\neq i}^{n-1} P(X_j | C, X_i) \quad (30)$$

The AODE classifier has been shown to be as accurate as TAN, but more computationally efficient in training. Also, since AODE performs model averaging rather than model selection, it has been shown to have lower variance [19].

Coming back to the problem of incorporating high-order feature interaction into MI-based feature selection, we face the similar problem of how to choose the feature X_j to condition on in Eq. (25). One possible approach would be to search for the optimal feature to condition on, similar to the TAN method for building relaxed Bayes classifiers. Our proposal of RelaxMRMR in this paper by averaging over all features is inspired by the AODE approach.

7. Conclusion

The range of MI-based feature selection approaches could be visualized as in Fig. 11. From left to right, the methods make use of increasingly higher-dimensional MI quantities and thus are able to detect increasingly higher-order feature dependencies. The associated cost is two-fold: (i) increased computational complexity, and (ii) larger amount of data is needed for accurate training.

From the left end, to our knowledge, there are MI-based methods that make use of feature dependency quantities up to second-order, for example the conditional relevancy $I(X_i;C|X_j)$ and joint mutual information $I(\{X_i,X_j\};C)$. From the right end, there are a few methods that make use of the full high-order dependency, i.e., the high-dimensional MI criterion $I(\{X_1,X_2,\ldots,X_m\};C)$ [21–23]. In-between second-order dependency and full high-order dependency, there is currently no or little research to our knowledge.

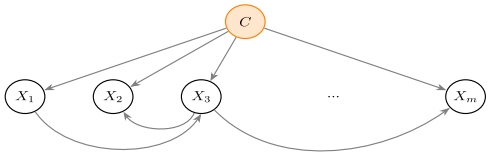

Fig. 9. Bayesian network representation of the Tree-Augmented Naive Bayes (TAN) classifier. Each node (feature) is allowed to have at most another parent apart from C.

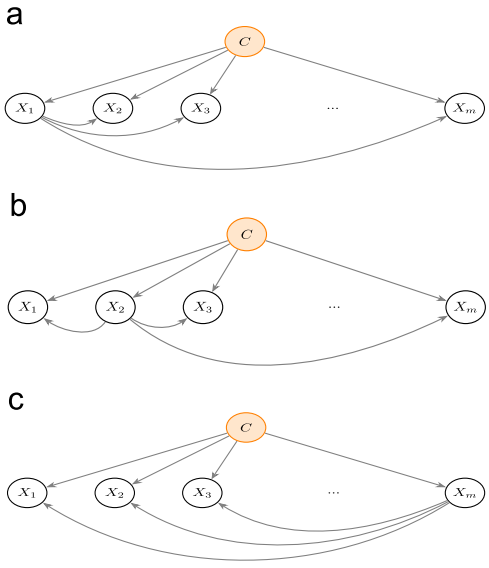

Fig. 10. Bayesian network representation of the base classifiers for the AODE model. Each feature takes turn to be the parent of all other features.

The theoretical framework presented in this paper hopes to stimulate more research to fill in this gap. We identified the assumptions needed for decomposing the full joint mutual information criterion into lower-dimensional MI quantities. We then proposed a principled approach for deriving new higher-dimensional MI based feature selection approaches by relaxing the identified assumptions. Our work is the first to explore the use of the three-way feature interaction terms $I(X_i;X_j|X_k)$. The proposed RelaxMRMR method is demonstrated to be effective via extensive experimental evaluation.

Conflict of interest

None declared.

Acknowledgements

This work is supported by the Australian Research Council via grant numbers FT110100112 and DP140101969.

References

[1] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[2] R. Fano, Transmission of Information: A Statistical Theory of Communications, MIT Press Classics, M.I.T. Press, 1961.
[3] G. Brown, A. Pocock, M.-J. Zhao, M. Luján, Conditional likelihood maximisation: a unifying framework for information theoretic feature selection, J. Mach. Learn. Res. 13 (2012) 27–66.
[4] R. Battiti, Using mutual information for selecting features in supervised neural net learning, IEEE Trans. Neural Netw. 5 (1994) 537–550.
[5] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 1226–1238.
[6] H.H. Yang, J. Moody, Detecting novel associations in large data sets, NIPS, 99, Citeseer, 1999 693-687.
[7] F. Fleuret, I. Guyon, Fast binary feature selection with conditional mutual information, J. Mach. Learn. Res. 5 (2004) 1531–1555.
[8] D. Lin, X. Tang, Conditional infomax learning: an integrated framework for feature extraction and fusion, in: ECCV'06, pp. 68–82.
[9] X.V. Nguyen, J. Chan, S. Romano, J. Bailey, Effective global approaches for mutual information based feature selection, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, 2014, pp. 512–521.
[10] K. Balagani, V. Phoha, On the feature selection criterion based on an approximation of multidimensional mutual information, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2010) 1342–1343.
[11] H. Cheng, Z. Qin, W. Qian, W. Liu, Conditional mutual information based feature selection, Knowledge Acquisition and Modeling, 2008. KAM'08. International Symposium on, IEEE, (2008) pp. 103–107.
[12] I. Rodriguez-Lujan, R. Huerta, C. Elkan, C.S. Cruz, Quadratic programming feature selection, J. Mach. Learn. Res. 11 (2010) 1491–1516.
[13] G. Herman, B. Zhang, Y. Wang, G. Ye, F. Chen, Mutual information-based method for selecting informative feature sets, Pattern Recognit. 46 (2013) 3315–3327.
[14] A. Frank, A. Asuncion, UCI machine learning repository, 2010.
[15] I. Guyon, S. Gunn, A. Ben-Hur, G. Dror, Result analysis of the nips 2003 feature selection challenge, Adv. Neural Inf. Process. Syst. (2003) 545–552.
[16] I. Rodriguez-Lujan, R. Huerta, C. Elkan, C.S. Cruz, Quadratic programming feature selection, J. Mach. Learn. Res. 11 (2010) 1491–1516.
[17] Z. Zhao, H. Liu, Spectral feature selection for supervised and unsupervised learning, Proceedings of the 24th International Conference on Machine Learning, ACM, (2007), pp. 1151–1157.

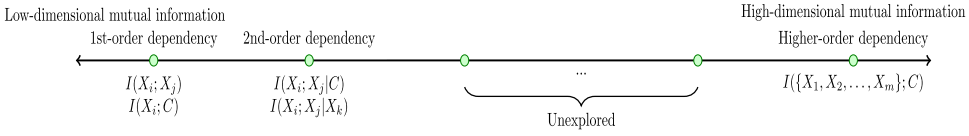

Fig. 11. A continuum of mutual information based feature selection methods.

[18] I. Kononenko, E. Šimec, M. Robnik-Šikonja, Overcoming the myopia of inductive learning algorithms with relief, Appl. Intell. 7 (1997) 39–55.
[19] G.I. Webb, J.R. Boughton, Z. Wang, Not so naive Bayes: aggregating one-dependence estimators, Mach. Learn. 58 (2005) 5–24.
[20] N. Friedman, D. Geiger, M. Goldszmidt, Bayesian network classifiers, Mach. Learn. 29 (1997) 131–163.
[21] Y. Zheng, C.K. Kwoh, A feature subset selection method based on high-dimensional mutual information, Entropy 13 (2011) 860–901.
[22] T.W.S. Chow, D. Huang, Estimating optimal feature subsets using efficient estimation of high-dimensional mutual information, Neural Networks, IEEE Transactions on 16 (2005) 213–224.
[23] N.X. Vinh, J. Chan, J. Bailey, Reconsidering mutual information based feature selection: A statistical significance view, The Twenty-Eighth AAAI Conference on Artificial Intelligence, American Association for Artificial Intelligence (AAAI) Press, (2014), pp. 2013–2019.

Nguyen Xuan Vinh is currently a research fellow at the University of Melbourne. He has made significant contribution to promoting information theoretic approaches for data mining, in particular clustering and feature selection, with more than 500 citations since 2010. His current research focuses on big data analytics in diverse domains, including bioinformatics, transportation and social networks.

Shuo Zhuo is a PhD candidate at the Department of Computing and Information Systems, the University of Melbourne. Prior to starting his research career, he completed a Master's degree in Distributed Computing at the University of Melbourne. His research focuses on tensor factorization approaches for big data.

Jeffrey Chan is a lecturer at RMIT University, Australia and an Honorary Research Fellow at the University of Melbourne. He has published over 35 papers in machine learning, graph analysis and social computing. He was previously a Research Fellow at the University of Melbourne, where he contributed to the paper.

James Bailey is a Professor and Australian Research Council (ARC) Future Fellow in the Department of Computing and Information Systems at the University of Melbourne. He has an extensive track record in databases and data mining and has been a chief investigator on multiple ARC discovery grants. He has been the recipient of five best paper awards and is an active member of the knowledge discovery community.