Pattern Recognition 67 (2017) 47–61

Contents lists available at ScienceDirect

Pattern Recognition

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

A new maximum relevance-minimum multicollinearity (MRmMC) method for feature selection and ranking

Azlyna Senawi[a], Hua-Liang Wei[a,b,*], Stephen A. Billings[a,b]

[a] Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield, S1 3JD, UK
[b] INSIGNEO Institute for in Silico Medicine, University of Sheffield, Sheffield, S1 3JD, UK

ARTICLE INFO

Article history:
Received 11 May 2016
Revised 9 December 2016
Accepted 18 January 2017
Available online 1 February 2017

Keywords:
Dimensionality reduction
Feature selection
Classification
Correlation measure
Qualitative and quantitative variables

ABSTRACT

A substantial amount of datasets stored for various applications are often high dimensional with redundant and irrelevant features. Processing and analysing data under such circumstances is time consuming and makes it difficult to obtain efficient predictive models. There is a strong need to carry out analyses for high dimensional data in some lower dimensions, and one approach to achieve this is through feature selection. This paper presents a new relevancy-redundancy approach, called the maximum relevance–minimum multicollinearity (MRmMC) method, for feature selection and ranking, which can overcome some shortcomings of existing criteria. In the proposed method, relevant features are measured by correlation characteristics based on conditional variance while redundancy elimination is achieved according to multiple correlation assessment using an orthogonal projection scheme. A series of experiments were conducted on eight datasets from the UCI Machine Learning Repository and results show that the proposed method performed reasonably well for feature subset selection.

1. Introduction

Technological advancement in data storage has led to the explosive growth in size of massive datasets which are usually of high dimensional with redundant and irrelevant features. Modelling high dimensional data is often computationally expensive and good predictive models are difficult to obtain because datasets may contain a large number redundant and irrelevant features. Thus, dimensionality reduction is seen as a crucial pre-processing step to overcome these problems which can be done by feature selection or feature extraction. In both approaches, the aim is to downscale a high dimensional data or feature space to a manageable low dimensional representation while retaining the data structure or useful information as much as possible.

In feature extraction approaches such as principal component analysis [1] and linear discriminant analysis [2], new features are constructed from the original features to form a new reduced dimensional space by combining or transforming the original features using some functional mapping. Although the new features in the new reduced dimensional space are related to the original features, the actual interpretation of the original features and hence the relation to the original system variables is completely lost in most cases. This drawback should be taken into account when considering dimensionality reduction since the actual interpretation may be important to understand the learning process that generates the new feature space [3]. Feature extraction also often associated with computational inefficiency despite the fact that it may significantly reduce dimensional space since the new constructed features are based on transformation that involves all original features including irrelevant and redundant features.

Unlike feature extraction which attempts to create new features based on all original features, feature selection is an approach which requires a selection of the most significant subset of features to a targeted concept by removing redundant and irrelevant features [4]. These redundant and irrelevant features can be ignored because they give very little or no unique information for data analysis and modelling.

Existing feature selection methods can be broadly categorized into three models: filter, wrapper and hybrid. Feature subset selection with a filter model is independent of specific mining algorithms as the search is based on the subset relevance to the targeted evaluation criterion. Hence, the filter model is not affected by any bias caused by the mining algorithm. The independent property also implies feature selection has to be carried out just once since the result can be used for different mining algorithms. In addition, the filter model is also considered as having simple search structure and is thus relatively easy to understand in comparison with other models.

* Corresponding author.
E-mail addresses: a.senawi@sheffield.ac.uk (A. Senawi), w.hualiang@sheffield.ac.uk (H.-L. Wei), s.billings@sheffield.ac.uk (S.A. Billings).

http://dx.doi.org/10.1016/j.patcog.2017.01.026
0031-3203/

In contrast to the filter model which selects feature subset relevant to the targeted evaluation criterion, the wrapper model selects a feature subset which is relevant to a predetermined mining algorithm. The mining algorithm is used as a black box to evaluate the quality of each candidate feature subset in order to find the best feature subset. This means that the wrapper model performs feature selection based on the mining performance level in which a feature subset is selected when the mining algorithm shows an optimal performance while taking into account feature dependencies in the feature selection procedure. As a result, the feature subset selected using the wrapper model will give higher mining performance than the filter model since the wrapper model is designed to search feature subset that is particularly tailored to the employed mining algorithm. For the same reason, however, rendering the feature subset obtained by the mining algorithm is unlikely to be suitable for use with other mining algorithms. Besides, the wrapper model is computationally slower when compared to the filter model since the mining algorithm of the wrapper model has to perform its task repeatedly until the final feature subset that gives maximum mining performance is found. This explains why the filter model is preferable than the wrapper model in handling large feature space problems.

The hybrid model emerged with an aim to combine the advantages possessed by both the filter and wrapper models. The model applies both an independent measure and a mining algorithm to measure the quality of each feature subset in the search space. Since mining performance is used as a guideline to stop the search, feature selection results based on the hybrid model is therefore specific to the mining algorithm employed. Consequently, the selected feature subset may not fit well with other mining algorithms and hence the hybrid model suffers the same problem as in the wrapper model.

Suppose that there are *M*original features in a dataset. An exhaustive search for an optimal feature subset when there exists 2^Mcandidate subsets is impractical for large M and even with a moderate M since it is too time consuming. Nevertheless, a search does not necessarily need to be exhaustive in order for it to be optimal as demonstrated in branch and bound method and best first search approaches. However, all optimal methods can be expected to be considerably slow for high dimensional problems [3]. Thus, it is often preferable for many high dimensional problems to employ heuristic methods that compromise subset optimality for better computational efficiency. A few examples of such search strategies are sequential search [5,6], floating search [7–9], random mutation hill climbing [10] and evolutionary-based approaches [11–14].

Much of the early work on feature selection focuses on choosing relevant features. Traditionally, feature redundancy was defined in some explicit or inexplicit manner, highlighting the need to remove redundant features [7,15–18]. Recently, a more concrete definition of feature redundancy was given in [19] with an illustration of the conceptual relationship between feature relevancy and redundancy. For example, in [17] the Markov Blanket filtering process was utilized to form the definition and an explicit redundancy analysis was also presented.

The concepts of feature relevancy and feature redundancy are translated and expressed by means of certain feature relationships in feature selection methods. The relevance of a feature is measured by evaluating its relationship with the target class label, while the redundancy of a feature is measured by its relationship with other features in the currently selected feature subset.

2. Related work

Many feature selection methods in the literature use mutual information to measure feature relevancy and redundancy. In [20], features are ranked according to their mutual information with respect to the class label and also with respect to the previously selected features. The mutual information based feature selection (MIFS) method proposed by Battiti [20] follows hill climbing selection scheme and chooses the next best feature that maximizes

$$J(\mathbf{f}_i) = I(\mathbf{c}, \mathbf{f}_i) - \beta \sum_{\mathbf{f}_j \in S} I(\mathbf{f}_j, \mathbf{f}_i) \tag{1}$$

where $I(\mathbf{c}, \mathbf{f}_i)$ denotes mutual information between class label $\mathbf{c}$ and candidate feature vector $\mathbf{f}_i$ while $I(\mathbf{f}_j, \mathbf{f}_i)$ denotes mutual information between previously selected feature $\mathbf{f}_j$ which have been accumulated in subset S and candidate feature $\mathbf{f}_i$. The parameter β is a user predefined value that will control the importance of redundant features. The larger the value, the more the measurement criterion will remove redundant features.

A variant of the MIFS method called the MIFS-U [21] emerged later to overcome the MIFS limitation which does not reflect relationships between feature and class label properly in its redundancy term if β is set too large. The MIFS-U approach brought a slight change to the right-hand side term so that the MIFS criterion becomes

$$J(\mathbf{f}_i) = I(\mathbf{c}, \mathbf{f}_i) - \beta \sum_{\mathbf{f}_j \in S} \frac{I(c, \mathbf{f}_j)}{H(\mathbf{f}_j)} I(\mathbf{f}_j, \mathbf{f}_i) \tag{2}$$

where $H(\mathbf{f}_j)$ is the entropy of $\mathbf{f}_j$. However, the MIFS-U approach is limited for uniformly distributed information.

As the number of features to be selected increases, the right-hand side term becomes incomparable with the left-hand side term for both MIFS and MIFS-U methods due to magnitude expansion of the right-hand side term [22]. Because of this problem, the methods may be forced to select and prioritize irrelevant features rather than relevant and/or redundant features. Another problem with both methods is that their optimal solution depends on the value assigned to β with optimal β's being considered subject to data structure. Hence no specific guided rule was given on how to choose parameter β. Apparently, a user may need to try different values before an optimal or acceptable suboptimal solution can be obtained.

The issue of incomparable terms in MIFS and MIFS-U methods mentioned earlier was overcome in the minimal-redundancy-maximum relevance (mRMR) feature selection criterion [23] by substituting β with reciprocal of the subset S cardinality, $1/|S|$. This will prevent the cumulative sum of the second term from having an excessive value in the expansion at any number of feature subsets to be considered which then lead to two equivalent terms for comparison. The mRMR criterion maximizes

$$J(\mathbf{f}_i) = I(\mathbf{c}, \mathbf{f}_i) - \frac{1}{|S|} \sum_{\mathbf{f}_j \in S} I(\mathbf{f}_j, \mathbf{f}_i). \tag{3}$$

Mutual information is preferable as an evaluation criterion over the correlation function for many proposed feature selection methods because of its ability to measure arbitrary dependence relationships between two features [20,24]. The method is not only limited to numerical features, but also applies to symbolic features consisting of discrete categories [24]. These two advantages made the mutual information based criterion to be seen as a more universal and robust measure.

Despite the aforementioned advantages, the mutual information criterion also has a few notable drawbacks. Mutual information computation is straightforward for discrete (categorical) random variables where an exact solution can be obtained easily. However, for continuous random variables which are frequently encountered in mutual information computations, it is difficult to gain the exact solution since the computation of the exact probability density functions (pdfs) is impossible [21]. Hence, an esti-

mation of the mutual information is required and different methods can be employed. Among the possible methods are histogram-based [25], kernel density estimation [26], k-nearest neighbour [27], Parzen window [28], B-spline [29], adaptive partitioning [30,31] and fuzzy-based [32] approaches. These estimation methods typically involve some pre-set parameters whose optimal values heavily depend on problem characteristics. Parameter settings could possibly be the major source of large estimation errors but still the parameters are often assigned with arbitrary values because there is no clear-cut rule provided [33]. In addition, there are so many available options for the mutual estimation calculations. Therefore, the efficiency of a feature selection approach greatly relies on the method applied.

In [34], another form of relevancy-redundancy measurement criterion similar to the three criteria discussed above (i.e., MIFS, MIFS-U and mRMR) was introduced particularly for continuous variables. This criterion, referred to as the *F*-test correlation difference (FCD), does not involve the calculation of mutual information. It selects the next best feature that maximizes

$$J(\mathbf{f}_i) = F(\mathbf{c}, \mathbf{f}_i) - \frac{1}{|S|}\sum_{\mathbf{f}_j \in S} \left| r(\mathbf{f}_j, \mathbf{f}_i) \right| \tag{4}$$

where $F(\mathbf{c}, \mathbf{f}_i)$ is the *F*-test statistic (or *t*-test statistic if two-class classification task is considered) comparing feature $\mathbf{f}_i$ and the class label $\mathbf{c}$ whereas $r(\mathbf{f}_j, \mathbf{f}_i)$ can be chosen to be Pearson correlation coefficient, Euclidean distance or any other appropriate measure. One problem with the FCD criterion is that the first term (*F*-test statistic) is not comparable with the second cluster of terms (redundancy terms) as they have different range of magnitude. The *F*-test statistic can take any positive value, while the value of redundancy coefficient ranging from zero to one. As a consequence, the *F*-test value may dominate the optimization criterion and reduce the impact of the second cluster of terms.

This paper presents a new alternative relevancy-redundancy criterion for feature selection, which is designed to take advantage of the idea of both the mRMR and FCD criteria, and meanwhile avoid the drawback of the two methods inherited from the original MIFS algorithm introduced in [20]. It is known that MIFS has a drawback in that its performance replies on the choice of the parameter beta for controlling and penalising the redundancy; the optimal choice of the parameter beta, however, strongly depends on the problem to be solved [22]. The proposed criterion is different from the two criteria in that it does not require any pre-specification or determination of thresholds for parameter settings. In the proposed method, relevant features are measured using conditional variance [35] while redundancy elimination is achieved through multiple correlation assessment using an orthogonal projection scheme [36]. The combination of these methods was motivated by the requirement to form a robust criterion that allow a comparable evaluation of feature relevancy and redundancy, yet avoiding mutual information based approach. Unlike mutual information based feature selection, the proposed method has the advantage of not demanding any control parameters, thus preventing any uncertainty associated with the method and providing consistency in the results.

The remaining contents of the paper are organized as follows. Section 3 is mainly reserved for a comprehensive discussion on how feature relevancy can be assessed by means of conditional correlation. Section 4 presents the idea of feature redundancy assessment by utilizing the concept of multicollinearity. The description also includes the interrelation of multicollinearity and squared multiple correlation coefficient, as well as how the coefficient can be used to quantify feature redundancy. A new feature selection criterion that tries to optimize both feature relevancy and feature redundancy is then introduced in Section 5. Section 6 gives details of the experimental setup and the procedure used in order to show the efficiency of the proposed method. The empirical results and extensive discussion are given in Section 7, followed by conclusion for the paper in Section 8.

3. Feature relevancy assessment

While many powerful feature selection methods were proposed in the literature to tackle various issues, relatively less and limited work has been done to assess the correlation between discrete (nominal) and continuous (quantitative) features directly. The majority of the prominent correlation measures were specifically designed for use either between two features of the same data type or between continuous and ordinal features.

The point-biserial correlation coefficient [37] is the most popular measure suggested when one feature is discrete while the other one is continuous. Yet the measure can only be used when the discrete feature is dichotomous or possibly be made dichotomous which is not always the case for many applications. An effort was made in [35] to fill this gap where a correlation measure between discrete and continuous features based on the underlying properties of marginal and conditional expectation and variance was introduced. The measure was adopted as part of the evaluation criterion for the feature selection approach that is specific to address some problem in mineral resources domain. In [38], an efficient correlation measure based filter (ECMBF) algorithm was proposed for the assessment of both feature relevancy and feature redundancy for more general applications. The ECMBF algorithm requires two predefined parameters, to distinguish weak irrelevance/relevance and redundancy, respectively. The choice of the two parameters can significantly affect the quality of the selected feature subset. This is probably the main disadvantage of the algorithm. Another drawback of ECBMF is that the assessment of the redundancy of each candidate feature is independent of the current selected features. In this study, an alternative approach is desired to overcome these drawbacks. The proposed correlation based method uses two measures that simultaneously evaluate features' dependency and redundancy, based on which 'best' features are selected using a sequential forward algorithm. The proposed method in this study is different from other types of filter approaches for example the Fisher score based methods [39].

In this paper, the potential of the correlation measure proposed in [35] is exploited; it will particularly be used to assess feature relevance. Towards better understanding the reliability of this correlation measure, its theoretical properties and conditions will be discussed first in detail.

Let X represent a quantitative random variable and Y represent a nominal random variable with some possible outcomes y_i. If every outcome y_i is described by a certain probability $P(Y = y_i)$ then the marginal expectation (also known as the expected value of X) symbolized by $E(X)$, is given by

$$E(X) = \sum_{y_i} P(Y = y_i)\, E(X|Y = y_i) \tag{5}$$

where $E(X|Y = y_i)$ denotes the conditional expectation of X given $Y = y_i$. It can be shown from this definition that the expected value of the conditional expectations, denoted by $E[E(X|Y)]$, is $E(X)$, that is

$$E(X) = E[E(X|Y)]. \tag{6}$$

Marginal variance of the random variable X is defined as

$$\mathrm{Var}(X) = E([X - E(X)]^2) = E(X^2) - [E(X)]^2. \tag{7}$$

Analogous to Eq. (7), the conditional variance of X given $Y = y_i$ is

$$\mathrm{Var}(X|Y) = E(X^2|Y) - [E(X|Y)]^2). \tag{8}$$

Note that Var($X|Y$) can be considered as a random variable, thereby theoretically permits the computation of its expected value

$$E[\mathrm{Var}(X|Y)] = E\{E(X^2|Y) - [E(X|Y)]^2\}. \quad (9)$$

Based on the additive law of expectation, the Eq. (9) can be rewritten as

$$E[\mathrm{Var}(X|Y)] = E[E(X^2|Y)] - E([E(X|Y)]^2). \quad (10)$$

Applying the relationship given by Eq. (6) to the first term at the right-hand side of Eq. (10) yields

$$E[\mathrm{Var}(X|Y)] = E(X^2) - E([E(X|Y)]^2). \quad (11)$$

Next, it is of interest to consider the variance of the conditional expectation, marked byVar[$E(X|Y)$]. Using the marginal variance definition given in (7), Var[$E(X|Y)$] can be expressed as

$$\mathrm{Var}[E(X|Y)] = E([E(X|Y)]^2) - [E(E(X|Y))]^2. \quad (12)$$

Applying (6) in Eq. (12) implies

$$\mathrm{Var}[E(X|Y)] = E([E(X|Y)]^2) - [E(X)]^2. \quad (13)$$

Then adding Eq. (11) to Eq. (13) gives

$$E[\mathrm{Var}(X|Y)] + \mathrm{Var}[\mathrm{E}(X|Y)] = E(X^2) - [E(X)]^2. \quad (14)$$

Notice that the right-hand side of Eq. (14) is equal to Var(X) as stated in (7). Hence, the following relationship is obtained

$$\mathrm{Var}(X) = E[\mathrm{Var}(X|Y)] + \mathrm{Var}[E(X|Y)] \quad (15)$$

which is well known as the law of total variance. A special case of the law is $\mathrm{Var}(X) = E[\mathrm{Var}(X|Y)] \Leftrightarrow \mathrm{Var}[E(X|Y)] = 0$. This biconditional implication is true when every conditional expectation given $Y = y_i$ is equal to the marginal expected value. Since variances can never be negative, it is apparent that $\mathrm{Var}(X) \geq E[\mathrm{Var}(X|Y)]$ and $\mathrm{Var}(X) \geq \mathrm{Var}[E(X|Y)]$.

From Eq. (15) it can be observed that the overall variability of a random variable Xconsists of two components. One component is the expected value of the conditional variance,$E[\mathrm{Var}(X|Y)]$, that quantifies the average variability within outcomes. Another component is the variance of the conditional means,Var[$E(X|Y)$], that indicates how much the variability is between outcomes. The former is considered in the correlation measure which will be presented next.

The correlation coefficient that measure the relationship between a quantitative random variable X and a nominal random variable Y is defined by

$$r_{\mathrm{qn}}(X,Y) = \left(1 - \frac{E[\mathrm{Var}(X|Y)]}{\mathrm{Var}(X)}\right)^{1/2} \quad (16)$$

which actually exploits the law of total variance. Based on previous discussions about Var(X) and Var[$E(X|Y)$], it can be verified that $0 \leq r_{\mathrm{qn}}(X, Y) \leq 1$. A value of $r_{\mathrm{qn}}(X, Y)$ approaching '1' indicates that there is a strong correlation or dependency between X and Y. Meanwhile, the value of $r_{\mathrm{qn}}(X, Y)$ approaching '0' suggests that there is a weak relationship between X and Y. If X and Y are totally independent or uncorrelated, then $r_{\mathrm{qn}}(X, Y) = 0$, which is the special case of the law of total variance mentioned before. On contrary, the presence of perfect dependency or correlation between X and Y is indicate by $r_{\mathrm{qn}}(X, Y) = 1$.

The above correlation coefficient will be used to measure feature relevance. It will be integrated with multiple correlation assessment in order to define a new feature selection criterion that can measure both feature relevancy and feature redundancy simultaneously. The multiple correlation assessment can be used to identify features with multicollinearity and thus can be used to detect and remove redundant features.

4. Multicollinearity redundancy and the squared multiple correlation coefficient

4.1. Multicollinearity redundancy

Assume that there are a total of Moriginal features in a dataset. Feature selection refers to a process of searching an optimal or suboptimal subset of m features from the M features [40]. The resulting feature subset from the process should essentially lead to a performance improvement or at least with minimal performance degradation as much as possible for the task under consideration. This objective can be realized by selecting representative features that hold important information characterizing all original features. In particular, it can be done by not only selecting features that have high relevancy to the targeted class but also have low redundancy within selected features.

An ultimate feature redundancy occurs if a feature has exact linear dependency with the current selected features and thus provides no extra information. While exact linear dependency is rarely present in many real data, a significant type of redundancy is also taken into account in such a way that features with any potential multicollinearity will be removed. Multicollinearity is a term to describe the presence of strong correlation or high linear dependency among two or more independent variables. This means that a feature with multicollinearity can be linearly estimated by a set of other features at some high level of accuracy and therefore suggests such a feature has redundant information. In comparison to features having ultimate redundancy, features with multicollinearity redundancy still provide some unique information but not important enough to give notable impact for effective data analysis tasks for example classification.

Multicollinearity can be identified from high values of the multiple correlation coefficient. However, since the actual interest is to assess predictive power of the current selected features in estimating a considered feature, the squared multiple correlation coefficient is often used instead of the multiple correlation coefficient. The squared multiple correlation coefficient specifically indicates the proportion of the variation in the considered feature that is predictable from the selected features. The value ranges from 0 to 1 with higher values implying a better predictive power. When a maximum value of the squared multiple correlation coefficient is obtained it indicates a full predictive power which is the ultimate redundancy. Thus, the ultimate redundancy can be regarded as the best achievable multicollinearity. Note that the squared multiple correlation coefficient can be computed by utilizing pairwise orthogonal projection of features already selected [4,41] . This will be further discussed in the next section.

4.2. The squared multiple correlation coefficient

Suppose that the set $F = \{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_M\}$ is a complete dataset of M features where each $\mathbf{f}_i = [f_1^{(i)}, f_2^{(i)}, \ldots, f_N^{(i)}]^{\mathrm{T}}$ is a feature vector composed by N observations. Also suppose that a subset S consisting $(k-1)$ features $\mathbf{f}_{i_1}, \mathbf{f}_{i_2}, \ldots, \mathbf{f}_{i_{k-1}}$ has already been selected from the set of Moriginal features. These $(k-1)$ features are then transformed into orthogonal variables $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_{k-1}$ using certain type of transformation. If the next feature $\mathbf{f} = \mathbf{f}_{i_k}$ is selected and included into S later on, then the kth orthogonal variable, $\mathbf{q}_k$, associated to $\mathbf{f}$ is calculated by

$$\mathbf{q}_k = \mathbf{f} - \frac{\mathbf{f}^{\mathrm{T}}\mathbf{q}_1}{\mathbf{q}_1^{\mathrm{T}}\mathbf{q}_1}\mathbf{q}_1 - \cdots - \frac{\mathbf{f}^{\mathrm{T}}\mathbf{q}_{k-1}}{\mathbf{q}_{k-1}^{\mathrm{T}}\mathbf{q}_{k-1}}\mathbf{q}_{k-1}. \quad (17)$$

The squared correlation coefficient between a feature $\mathbf{f} \in F - S$ and an orthogonal variable $\mathbf{q} \in \{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_k\}$ is defined as

$$\mathrm{sc}(\mathbf{f}, \mathbf{q}) = \frac{(\mathbf{f}^{\mathrm{T}}\mathbf{q})^2}{(\mathbf{f}^{\mathrm{T}}\mathbf{f})(\mathbf{q}^{T}\mathbf{q})} = \frac{\left(\sum_{i=1}^{N} f_i q_i\right)^2}{\sum_{i=1}^{N} f_i^2 \sum_{i=1}^{N} q_i^2}. \quad (18)$$

Based on (18), the squared multiple correlation coefficient for each remaining feature $\mathbf{f} \in F - S$ with the selected features $\mathbf{f}_{i_1}, \mathbf{f}_{i_2}, \ldots, \mathbf{f}_{i_k}$ (or equivalently with $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_k$) can be computed as

$$R^2(\mathbf{f}; \mathbf{q}_1, \ldots, \mathbf{q}_k) = \sum_{\mathrm{i}=1}^{k} \mathrm{sc}(\mathbf{f}, \mathbf{q}_i) \quad (19)$$

where the square root of R^2 geometrically represents the length of orthogonal projection of $\mathbf{f}$ in the directions of the orthogonal variables $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_k$ divided by the norm (energy) of $\mathbf{f}$.

5. Monitoring criterion

In order to choose features that are most relevant to the targeted class $\mathbf{c}$, the monitoring condition is to maximize the measure V:

$$V = r_{\mathrm{qn}}^2(\mathbf{f}_j, \mathbf{c}) \quad \text{suchthat} \quad \mathbf{f}_j \in F - S \quad (20)$$

which utilizes the squared value of the correlation coefficient given in (16). On the other hand, the squared multiple correlation coefficient defined in (19) is suggested to guide selection of features that are least mutually dissimilar or least redundant. Thus, the redundancy condition to be considered for measuring redundancy between feature $\mathbf{f}_j$ and the current selected feature subset S is to minimize the measure W:

$$W = R^2(\mathbf{f}_j; \mathbf{q}_1, \ldots, \mathbf{q}_k) = \sum_{\mathrm{i}=1}^{k} \mathrm{sc}(\mathbf{f}_j, \mathbf{q}_i) \quad \text{suchthat} \quad \mathbf{f}_j \in F - S \quad (21)$$

where $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_k$ are orthogonal variables associated respectively with preceding selected features $\mathbf{f}_{i_1}, \mathbf{f}_{i_2}, \ldots, \mathbf{f}_{i_k}$ contained in S.

Because the aim of the feature selection is to select features that are highly relevant to the targeted class $\mathbf{c}$ and also has low redundancy with other selected features, both measures V and W are optimized simultaneously. A new feature to be added will be based on one possible single criterion combining both measures. The monitoring criterion used in this study is to maximize

$$J(\mathbf{f}_j) = r_{\mathrm{qn}}^2(\mathbf{f}_j, \mathbf{c}) - R^2(\mathbf{f}_j; \mathbf{q}_1, \ldots, \mathbf{q}_k) \quad \text{such that} \quad \mathbf{f}_j \in F - S \quad (22)$$

which can also be written as

$$J(\mathbf{f}_j) = \max_{\mathbf{f}_j \in F-S} \left[r_{\mathrm{qn}}^2(\mathbf{f}_j, \mathbf{c}) - \sum_{\mathrm{i}=1}^{k} \mathrm{sc}(\mathbf{f}_j, \mathbf{q}_i) \right]. \quad (23)$$

The correlation coefficient r_{qn} is squared in (22) so that a fair comparison can be made with the R^2 term. Clearly, there is no pre-defined parameter required from user in the criterion. The feature selection method, based on the criterion (23), is referred to as the maximum relevance – minimum multicollinearity (MRmMC) method.

In the MRmMC method, the first feature is selected if it satisfies the optimization criteria stated in (20) and the rest are selected based on criterion (23) by using forward sequential search strategy. At every subsequent step, a new feature will be added to previously selected feature subset. This simple piecewise feature search strategy will avoid excessive computational burden to the MRmMC feature selection, and can therefore accelerate the feature search procedure. Note that although the search may lead to a suboptimal solution, it can meet the requirements for most real applications.

Table 1
A summary of the datasets characteristics.

Dataset	Number of features	Number of instances	Number of classes
Glass [N]	9	214	7
Magic Gamma [N]	10	19,020	2
Vowel [N]	10	990	11
Statlog [N]	18	846	4
Mfeat Zernike [N]	47	2000	10
Sonar	60	208	2
Musk [N]	166	476	2
Mfeat Factors [N]	216	2000	10

[N]: The raw dataset was normalized for the proposed method in the experiment. This also means the dataset was normalized in classification accuracy computation for all classifiers.

The proposed criterion (22) can overcome the drawback of the MIFS approach, and it can effectively manage relevance and redundancy as follows. The first part, V, measures relevance using a correlation coefficient defined by (16) and (20), while the second part, W, measures the redundancy of a candidate feature with features in a selected feature set by evaluating the multicollinearity when the candidate feature is added to the existing feature subset.

The proposed criterion has the following advantages: i) The two parts of the criterion are comparable, and can result in a good balance between relevance and redundancy; ii) There is no need to pre-specify a control parameter as required in MIFS, and iii) the algorithm is relatively easier to implement. Some implementation details (pseudo-code) of MRmMC is shown in Fig. 1.

The time complexity of the MRmMC method is determined by three main parts: the assessment of feature relevancy to the class label, the computation of the squared correlation coefficient, and the orthogonalization operations. Feature relevancy assessment has a linear time complexity of $O(MN)$, where $M$is the number of candidate features and N is the number of observations. The computation of the squared correlation coefficient has a worst-case time complexity of $O(M^2N)$ while the orthogonalisation procedure is of a complexity of $O((M-1)N)$. As a result, the overall time complexity takes the order of $O(M^2N)$.

6. Experimental setup and procedure

A series of experiments were conducted to test and analyse the efficacy of the proposed MRmMC method from several perspectives. Eight datasets were used as benchmarks, and relevant results were compared with those generated from mRMR and MIFS.

6.1. Benchmark datasets

The eight public real datasets available from the UCI Machine Learning Repository, are depicted in Table 1. In order to provide comprehensive evaluation, the datasets were picked based on three different categories of dimensional size: low-dimension ($M \leq 10$), medium-dimension ($10 < M \leq 100$), and high-dimension ($M > 100$). Important details of the chosen datasets are summarized in Table 1. Observe that the datasets are also varied in terms of number of instances and number of classes.

6.2. Comparison with similar methods

The MIFS and mRMR methods are specifically employed for a comparison purpose as they possess similar forms of measurement criteria and use the same sequential feature search strategy. Feature subset solutions of the MIFS and mRMR methods were obtained by running the Feature Selection Toolbox (FEAST) (available

Input: $F = \{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_M\}$ // A complete dataset of M features
Output: S // Subset of features
Initialize: $L_1 = \{1, 2, \ldots, M\}$, $S = \{\}$
m // Number of features to be selected
for $j = 1$ to M
Compute $V_j = r_{qn}^2(\mathbf{f}_j, \mathbf{c})$ such that $\mathbf{f}_j \in F$;
end for
$l_1 = \arg\max_{j \in L_1} [V_j]$ such that $l_1 \in L_1$; $\mathbf{q}_1 = \mathbf{f}_{l_1}$; $\mathbf{z}_1 = \mathbf{f}_{l_1}$;
add $\mathbf{z}_1$ to S;
for $h = 2$ to m
$L_h = L_{h-1} \setminus \{l_{h-1}\}$; $k = h - 1$;
for $j \in H_l$
Find $J(\mathbf{f}_j) = r_{qn}^2(\mathbf{f}_j, \mathbf{c}) - \sum_{i=1}^{k} \mathrm{sc}(\mathbf{f}_j, \mathbf{q}_i)$;
end for
$l_h = \arg\max_{\mathbf{f}_j \in F - S} [J(\mathbf{f}_j)]$ such that $l_h \in L_h$
$\mathbf{q}_h = \mathbf{f}_{l_h} - \frac{\mathbf{f}_{l_h}^{\mathrm{T}} \mathbf{q}_1}{\mathbf{q}_1^{\mathrm{T}} \mathbf{q}_1} \mathbf{q}_1 - \cdots - \frac{\mathbf{f}_{l_h}^{\mathrm{T}} \mathbf{q}_{h-1}}{\mathbf{q}_{h-1}^{\mathrm{T}} \mathbf{q}_{h-1}} \mathbf{q}_{h-1}$;
$\mathbf{z}_h = \mathbf{f}_{l_h}$;
add $\mathbf{z}_h$ to S
end for

Fig. 1. The MRmMC algorithm.

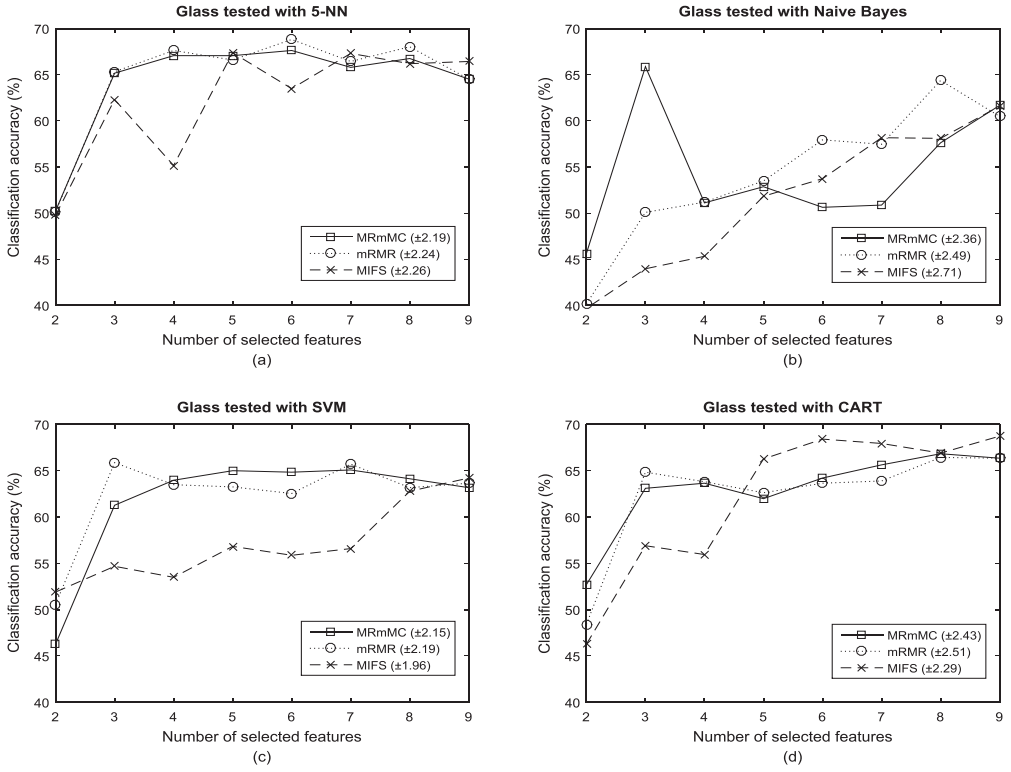

Fig. 2. Classification results for Glass dataset over different number of selected features, tested with four classifiers: (a) 5-NN, (b) Naïve Bayes, (c) SVM and (d) CART. Each plot shows comparison among MRmMC, mRMR and MIFS methods.

at: http://www.cs.man.ac.uk/~gbrown/fstoolbox/) that was originally developed by Brown et al. [42]. In this work, the redundancy parameter was chosen to be $\beta = 1$ for the MIFS method. This choice of parameter value was in the appropriate range suggested by Battiti [20].

6.3. Validation classifiers

MRmMC is a filter method, and hence its efficiency might be different from one classifier to another classifier. Thus, four of the ten most influential algorithms in data mining [43], namely, the

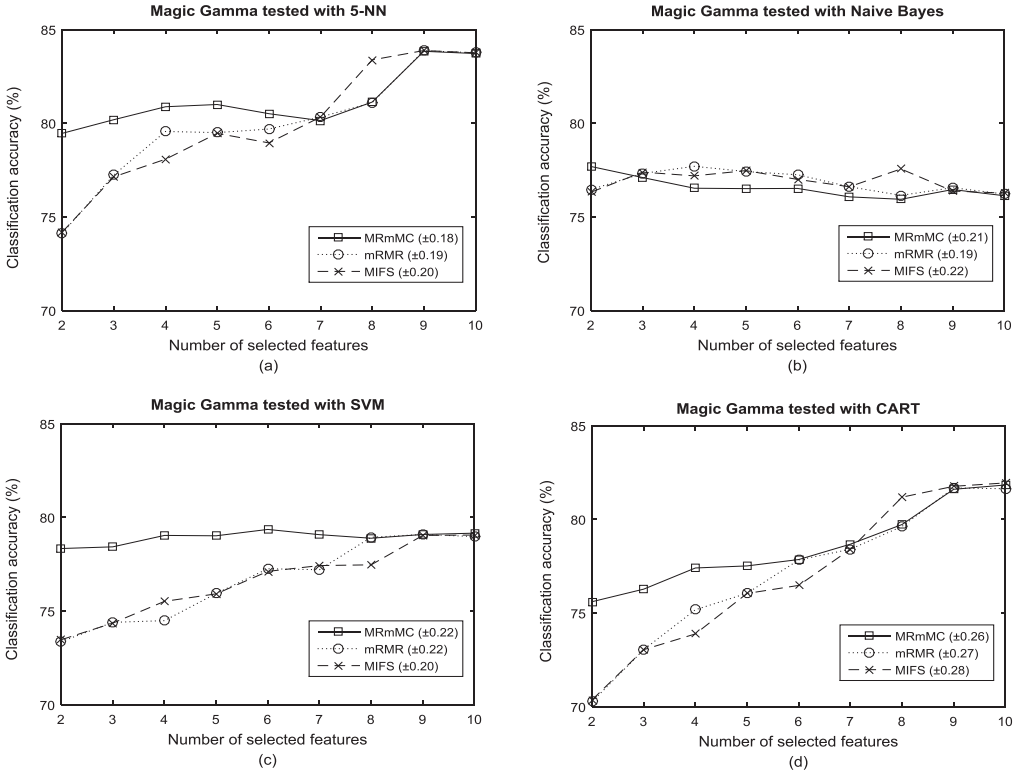

Fig. 3. Classification results for Magic Gamma dataset over different number of selected features, tested with four classifiers: (a) 5-NN, (b) Naïve Bayes, (c) SVM and (d) CART. Each plot shows comparison among MRmMC, mRMR and MIFS methods.

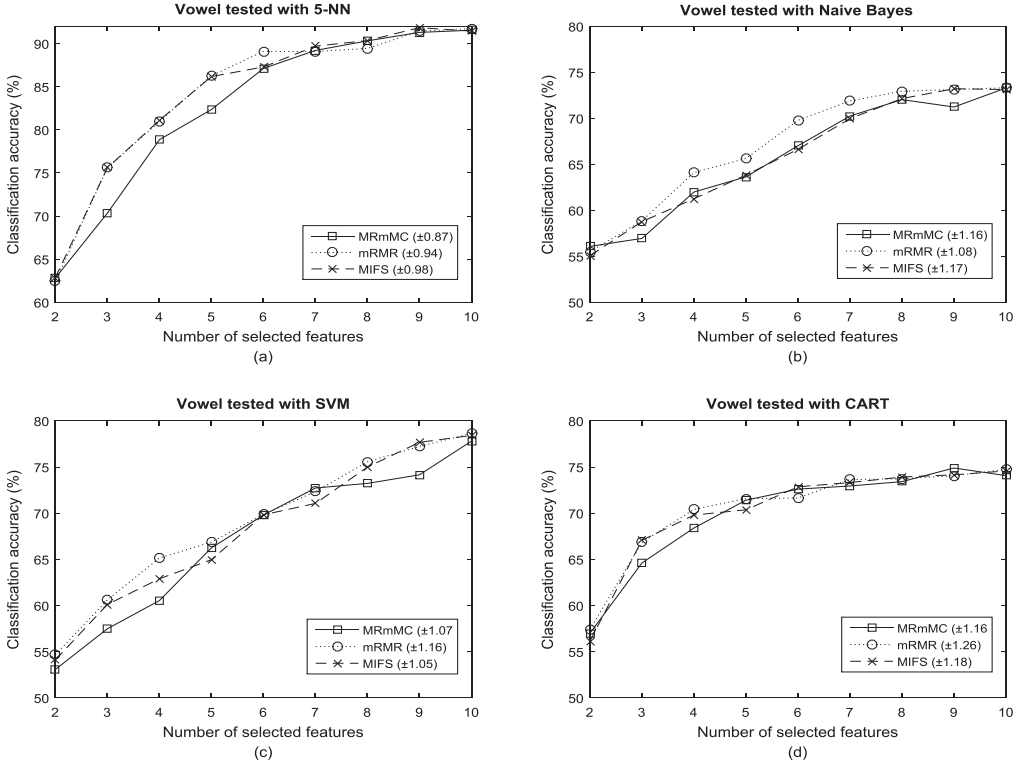

Fig. 4. Classification results for Vowel dataset over different number of selected features, tested with four classifiers: (a) 5-NN, (b) Naïve Bayes, (c) SVM and (d) CART. Each plot shows comparison among MRmMC, mRMR and MIFS methods.

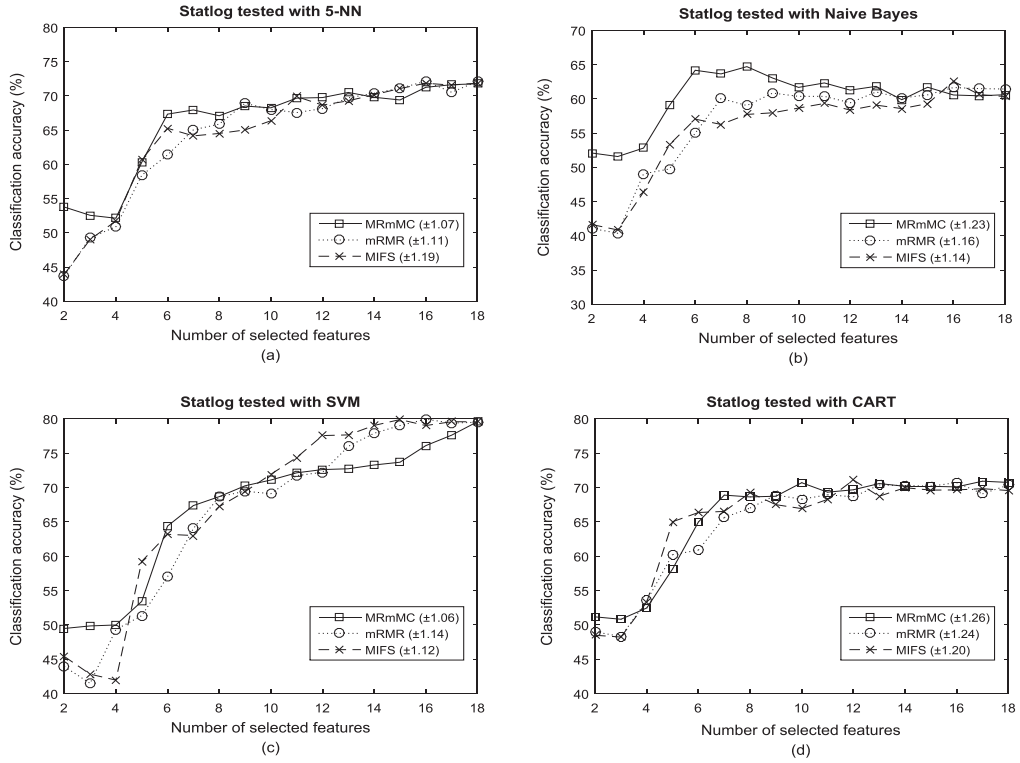

Fig. 5. Classification results for Statlog dataset over different number of selected features, tested with four classifiers: (a) 5-NN, (b) Naïve Bayes, (c) SVM and (d) CART. Each plot shows comparison among MRmMC, mRMR and MIFS methods.

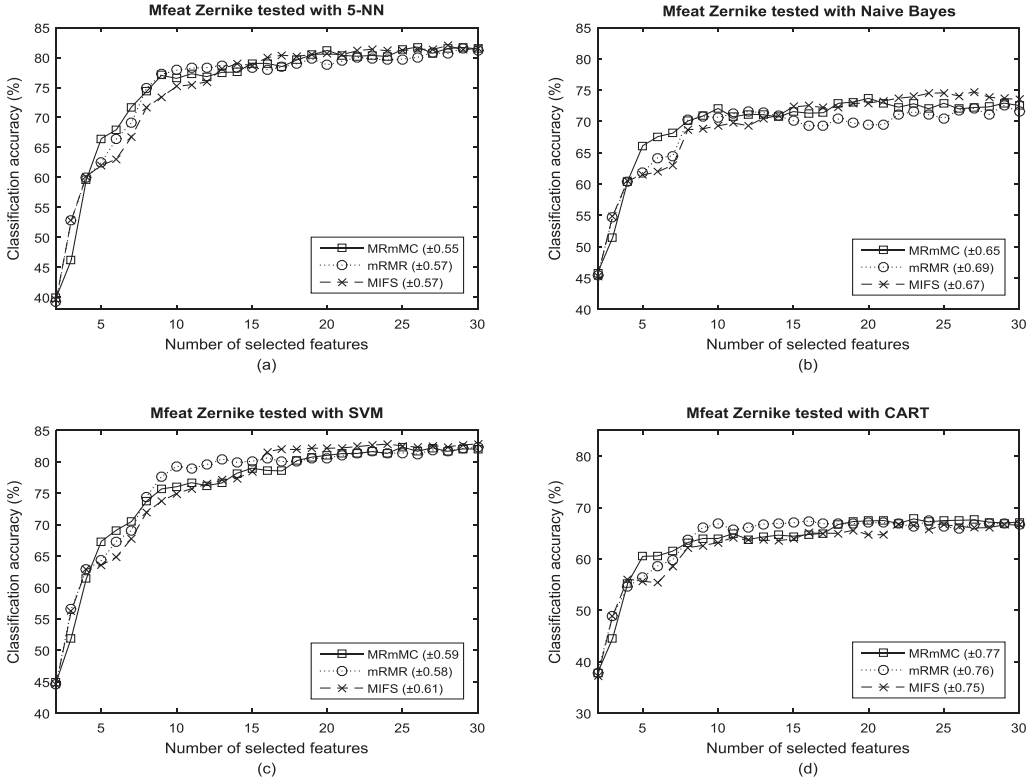

Fig. 6. Classification results for Mfeat Zernike dataset over different number of selected features, tested with four classifiers: (a) 5-NN, (b) Naïve Bayes, (c) SVM and (d) CART. Each plot shows comparison among MRmMC, mRMR and MIFS methods.

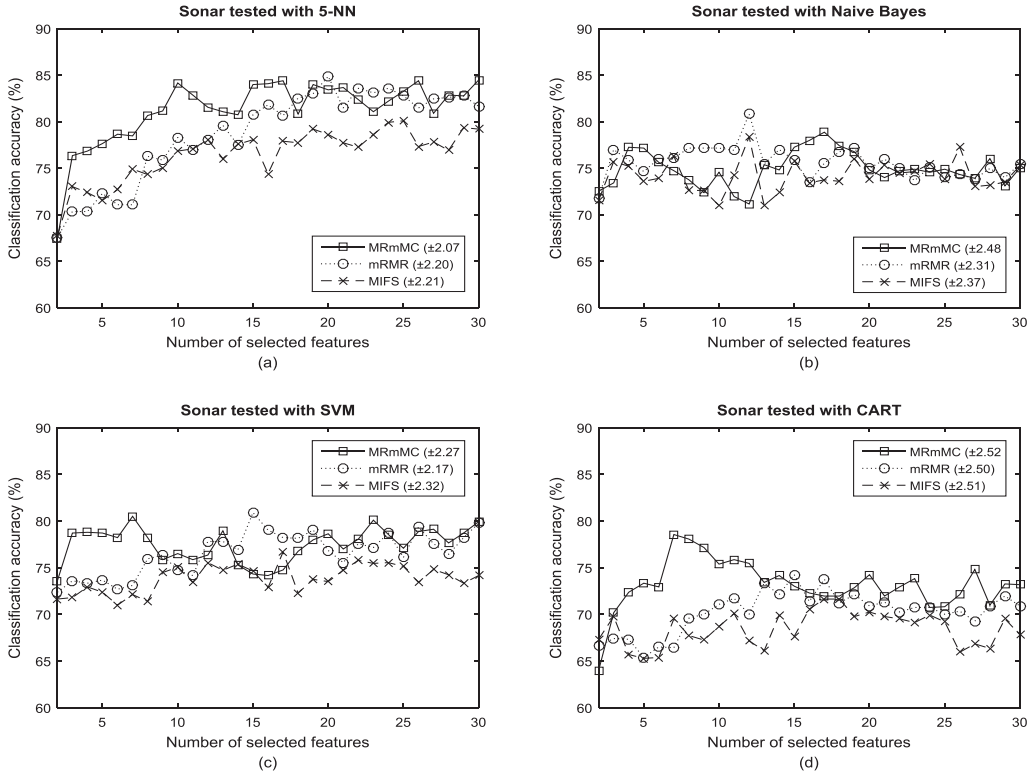

Fig. 7. Classification results for Sonar dataset over different number of selected features, tested with four classifiers: (a) 5-NN, (b) Naïve Bayes, (c) SVM and (d) CART. Each plot shows comparison among MRmMC, mRMR and MIFS methods.

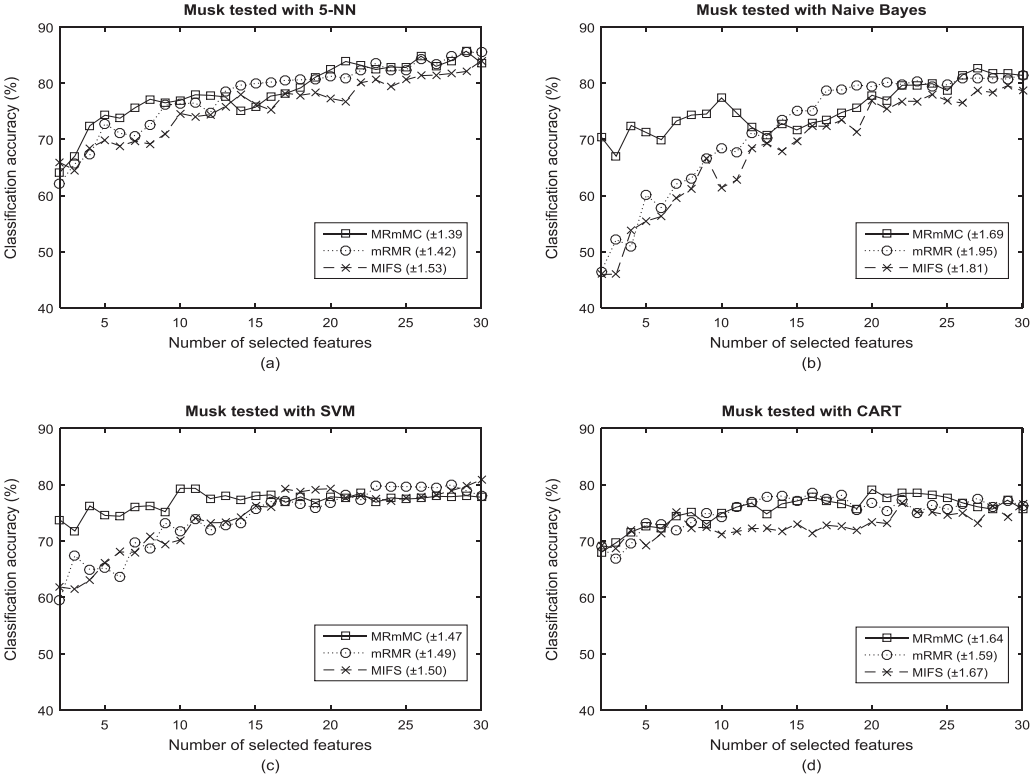

Fig. 8. Classification results for Musk dataset over different number of selected features, tested with four classifiers: (a) 5-NN, (b) Naïve Bayes, (c) SVM and (d) CART. Each plot shows comparison among MRmMC, mRMR and MIFS methods.

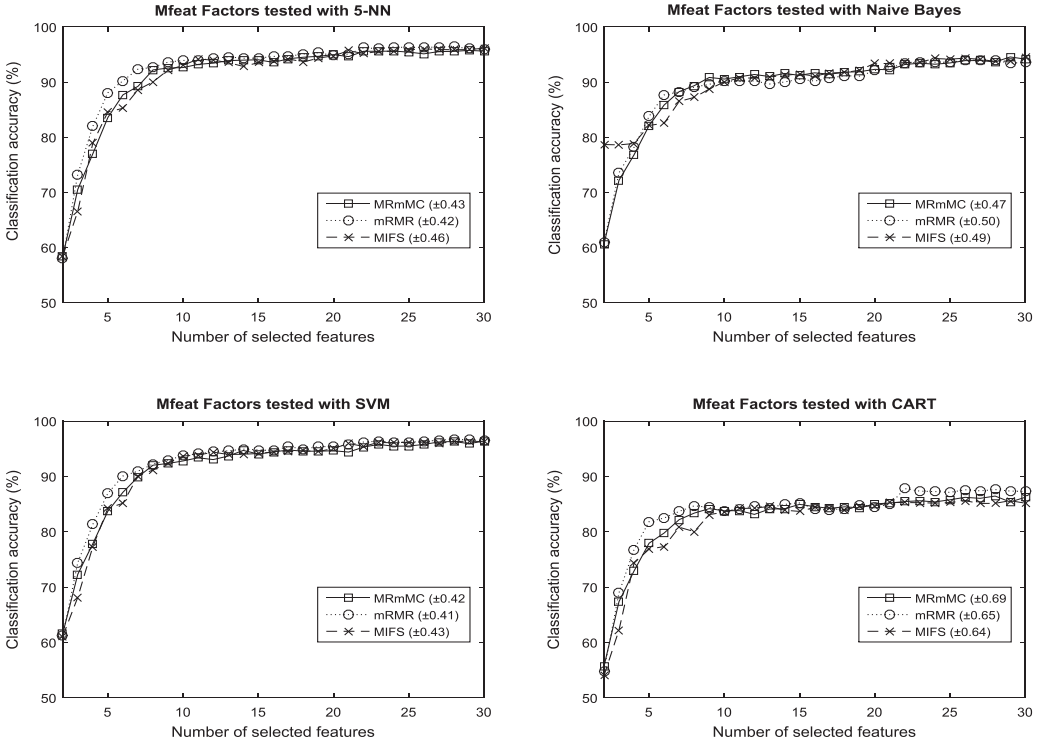

Fig. 9. Classification results for Mfeat Factors dataset over different number of selected features, tested with four classifiers: (a) 5-NN, (b) Naïve Bayes, (c) SVM and (d) CART. Each plot shows comparison among MRmMC, mRMR and MIFS methods.

Table 2
A comparison of the average classification accuracy based on the first *m* selected features.

	Glass					Magic Gamma				
	MRmMC	mRMR		MIFS		MRmMC	mRMR		MIFS	
5-NN	Accuracy	Accuracy	*p*-value	Accuracy	*p*-value	Accuracy	Accuracy	*p*-value	Accuracy	*p*-value
m = 5	62.38	62.42	0.51	58.65	0.01 *	80.38	77.61	0.00 *	77.22	0.00 *
m = 10	64.28	64.68	0.60	62.25	0.10	81.21	79.91	0.00 *	79.91	0.00 *
N Bayes	Accuracy	Accuracy	*p*-value	Accuracy	*p*-value	Accuracy	Accuracy	*p*-value	Accuracy	*p*-value
m = 5	53.87	48.73	0.00 *	45.20	0.00 *	76.96	77.22	0.96 ■	77.09	0.79
m = 10	54.53	54.40	0.47	51.55	0.05	76.55	76.85	0.98 ■	76.91	0.99 ■
SVM	Accuracy	Accuracy	*p*-value	Accuracy	*p*-value	Accuracy	Accuracy	*p*-value	Accuracy	*p*-value
m = 5	59.13	60.79	0.87	54.22	0.00 *	78.71	74.55	0.00 *	74.82	0.00 *
m = 10	61.72	62.28	0.64	57.04	0.00 *	78.93	76.63	0.00 *	76.60	0.00 *
CART	Accuracy	Accuracy	*p*-value	Accuracy	*p*-value	Accuracy	Accuracy	*p*-value	Accuracy	*p*-value
m = 5	60.36	59.92	0.40	56.35	0.01 *	76.70	73.64	0.00 *	73.34	0.00 *
m = 10	63.06	62.5	0.38	62.17	0.30	78.50	77.08	0.00 *	77.02	0.00 *
	Vowel					Statlog				
	MRmMC	mRMR		MIFS		MRmMC	mRMR		MIFS	
5-NN	Accuracy	Accuracy	*p*-value	Accuracy	*p*-value	Accuracy	Accuracy	*p*-value	Accuracy	*p*-value
m = 5	73.6	76.32	1.00 ■	76.45	1.00 ■	54.69	50.57	0.00 *	51.34	0.00 *
m = 10	82.66	84.01	0.98 ■	84.05	0.98 ■	61.99	59.06	0.00 *	58.97	0.00 *
m = 15	–	–	–	–	–	64.79	62.75	0.01 *	62.84	0.01 *
m = 30	–	–	–	–	–	65.99	64.31	0.02 *	64.42	0.03 *
N Bayes	Accuracy	Accuracy	*p*-value	Accuracy	*p*-value	Accuracy	Accuracy	*p*-value	Accuracy	*p*-value
m = 5	59.67	61.03	0.96 ■	59.73	0.53	53.88	45.06	0.00 *	45.55	0.00 *
m = 10	65.83	67.24	0.96 ■	66	0.58	59.20	52.84	0.00 *	52.21	0.00 *
m = 15	–	–	–	–		59.99	55.51	0.00 *	54.61	0.00 *
m = 30	–	–	–	–		60.08	56.57	0.00 *	55.77	0.00 *
SVM	Accuracy	Accuracy	*p*-value	Accuracy	*p*-value	Accuracy	Accuracy	*p*-value	Accuracy	*p*-value
m = 5	59.34	61.83	1.00 ■	60.53	0.94	50.7	46.54	0.00 *	47.37	0.00 *
m = 10	67.23	69.00	0.99 ■	68.23	0.90	60.51	57.16	0.00 *	58.25	0.00 *
m = 15	–	–	–	–	–	64.93	63.67	0.06	65.20	0.63
m = 30	–	–	–	–	–	67.2	66.48	0.18	67.71	0.74
CART	Accuracy	Accuracy	*p*-value	Accuracy	*p*-value	Accuracy	Accuracy	*p*-value	Accuracy	*p*-value
m = 5	65.35	66.56	0.92	65.84	0.72	53.16	52.78	0.34	53.77	0.75
m = 10	69.93	70.45	0.72	70.25	0.65	61.62	60.21	0.06	61.30	0.36
m = 15	–	–	–	–	–	64.61	63.60	0.13	64.25	0.34
m = 30	–	–	–	–	–	65.67	64.74	0.15	65.21	0.30

Table 3
A comparison of the average classification accuracy based on the first m selected features.

	Mfeat Zernike					Sonar				
	MRmMC	mRMR		MIFS		MRmMC	mRMR		MIFS	
5-NN	Accuracy	Accuracy	*p*-value	Accuracy	*p*-value	Accuracy	Accuracy	*p*-value	Accuracy	*p*-value
m = 5	53.06	53.66	0.90	53.64	0.88	74.55	70.13	0.00 *	71.16	0.02 *
m = 10	64.43	64.46	0.53	62.74	0.00 *	77.92	72.56	0.00 *	73.15	0.00 *
m = 15	69.15	69.42	0.73	67.98	0.00 *	79.39	74.7	0.00 *	74.65	0.00 *
m = 30	75.05	74.78	0.25	74.70	0.19	81.24	78.76	0.05	76.45	0.00 *
N Bayes	Accuracy	Accuracy	*p*-value	Accuracy	*p*-value	Accuracy	Accuracy	*p*-value	Accuracy	*p*-value
m = 5	55.96	55.58	0.24	55.54	0.20	75.08	74.81	0.43	74.00	0.27
m = 10	63.62	62.52	0.02 *	61.55	0.00 *	74.59	75.87	0.78	73.59	0.28
m = 15	66.28	65.57	0.08	64.77	0.00 *	74.41	76.35	0.88	73.86	0.37
m = 30	69.5	68.24	0.00 *	69.30	0.34	74.93	75.62	0.66	74.15	0.33
SVM	Accuracy	Accuracy	*p*-value	Accuracy	*p*-value	Accuracy	Accuracy	*p*-value	Accuracy	*p*-value
m = 5	56.4	57.08	0.88	56.82	0.78	77.44	73.23	0.01 *	72.18	0.00 *
m = 10	65.63	66.24	0.88	64.51	0.01 *	77.67	73.97	0.01 *	72.52	0.00 *
m = 15	69.81	71.08	1.00 ■	68.97	0.04 *	77.12	75.23	0.12	73.31	0.01 *
m = 30	75.66	76.31	0.94	75.89	0.70	77.48	76.58	0.29	73.86	0.01 *
CART	Accuracy	Accuracy	*p*-value	Accuracy	*p*-value	Accuracy	Accuracy	*p*-value	Accuracy	*p*-value
m = 5	49.54	49.47	0.45	49.45	0.44	69.96	66.67	0.04 *	67.01	0.07
m = 10	56.83	57.00	0.62	55.51	0.01 *	73.54	67.81	0.00 *	67.4	0.00 *
m = 15	59.53	60.40	0.94	58.46	0.03 *	73.84	69.4	0.01 *	67.68	0.00 *
m = 30	63.37	63.71	0.73	62.27	0.02 *	73.16	70.25	0.05	68.46	0.00 *
	Musk					Mfeat Factors				
	MRmMC	mRMR		MIFS		MRmMC	mRMR		MIFS	
5-NN	Accuracy	Accuracy	*p*-value	Accuracy	*p*-value	Accuracy	Accuracy	*p*-value	Accuracy	*p*-value
m = 5	69.49	66.98	0.02 *	67.18	0.02 *	72.36	75.33	1.00 ■	72.13	0.32
m = 10	73.12	70.52	0.01 *	69.12	0.00 *	82.63	84.90	1.00 ■	81.95	0.05
m = 15	74.45	73.16	0.12	71.48	0.00 *	86.59	88.25	1.00 ■	86.11	0.10
m = 30	78.53	78.02	0.31	75.72	0.00 *	90.98	92.10	1.00 ■	90.82	0.31
N Bayes	Accuracy	Accuracy	*p*-value	Accuracy	*p*-value	Accuracy	Accuracy	*p*-value	Accuracy	*p*-value
m = 5	70.3	52.41	0.00 *	50.31	0.00 *	72.91	74.09	0.99 ■	79.56	1.00 ■
m = 10	72.3	58.61	0.00 *	56.26	0.00 *	81.83	82.35	0.90	83.69	1.00 ■
m = 15	72.35	63.24	0.00 *	60.33	0.00 *	85.18	85.11	0.43	86.31	1.00 ■
m = 30	75.58	71.78	0.00 *	68.51	0.00 *	89.22	89.05	0.31	89.92	0.98 ■
SVM	Accuracy	Accuracy	*p*-value	Accuracy	*p*-value	Accuracy	Accuracy	*p*-value	Accuracy	*p*-value
m = 5	74.09	64.29	0.00 *	63.14	0.00 *	73.85	75.96	1.00 ■	72.69	0.01 *
m = 10	75.31	67.14	0.00 *	66.58	0.00 *	83.28	84.86	1.00 ■	82.57	0.04 *
m = 15	76.29	69.42	0.00 *	69.31	0.00 *	87.02	88.33	1.00 ■	86.68	0.18
m = 30	77.01	74.00	0.00 *	74.02	0.00 *	91.32	92.26	1.00 ■	91.29	0.46
CART	Accuracy	Accuracy	*p*-value	Accuracy	*p*-value	Accuracy	Accuracy	*p*-value	Accuracy	*p*-value
m = 5	70.51	69.64	0.22	69.75	0.25	68.45	70.60	1.00 ■	66.84	0.00 *
m = 10	72.43	71.78	0.29	71.27	0.16	76.34	77.93	1.00 ■	74.68	0.00 *
m = 15	73.80	73.72	0.47	71.61	0.03 *	79.08	80.33	0.99 ■	78.05	0.02 *
m = 30	75.59	75.25	0.39	72.93	0.01 *	82.32	83.37	0.99 ■	81.64	0.08

k-Nearest Neighbour (*k*-NN), Naïve Bayes, Support Vector Machine (SVM) and CART classifier algorithms, are used to verify the classification capability of the performance of the MRmMC method for feature subset selection. These classifiers were chosen not only because of their popularity but also because of their distinct learning mechanism. The aim is to test the overall performance of the newly proposed method in comparison to these popular classifiers.

Note that the number of nearest neighbours in the *k*NN classifier was chosen to be $k = 5$ in all experiments, and this is a fair choice for all the three methods: MRmMC, mRMR and MIFS.

6.4. Cross validation procedure

For each of the classifiers, a same holdout cross-validation scheme was used to test the performance. Particularly, 80% of the data were used for training whereas the remaining 20% were holdout (for testing) and once the training completed, these holdout data were then used to assess the spotted classification models in the testing stage.

In addition, to reduce variability in the assessment, 30 rounds of cross-validation were performed. The validation results are presented as the 95% confidence intervals for the classification accuracies based on the accuracies obtained from that 30 rounds.

7. Numerical results and discussion

Figs. 2–9 show classification results over different number of selected features by the three feature selection methods, tested with the four classifiers. The x-axis in each figure represents the number of selected features while the y-axis represents the average classification accuracy based on 30 rounds of cross-validation. For clear visualization and due to space limitations, the plots only present the performance of the first 30 selected features even if more than 30 were selected. This doesn't affect the performance evaluation of the feature selection methods.

It can be observed that the overall pattern of the classification accuracies of the three methods based on the selected feature subset for Mfeat Zernike and Mfeat Factors datasets is comparable to each other for all the four classifiers as illustrated in Figs. 6 and 9, respectively. Interestingly, the classification accuracy by MRmMC outperforms the other two methods if only a few number of significant features need to be identified, and as more features were progressively added, MRmMC gains the same level of accuracy as the other two. This pattern is particularly distinct for Magic Gamma, Vowel, Statlog, Mfeat Zernike, Sonar and Musk datasets as depicted in Figs. 3, 4, 5, 6, 7 and 8, respectively.

Table 4
The least number of selected features, m_{least}, by MRmMC, mRMR and MIFS methods that gives classification accuracy close to (at most 5% less than the full set accuracy) or better than the full feature set. The symbol "•" (or "□") denotes the proposed method has lower (or larger) value of m_{least} than the compared method. Results are based on Glass, Magic Gamma, Vowel, Statlog, Mfeat Zernike and Sonar datasets.

	Glass			Magic Gamma		
5-NN	Full set accuracy	m_{least}	Subset accuracy	Full set accuracy	m_{least}	Subset accuracy
MRmMC	64.52 ± 2.61	3	65.16 ± 1.97	83.72 ± 0.16	2	79.46 ± 0.18
mRMR	64.52 ± 1.96	3	65.32 ± 1.86	83.76 ± 0.20	4 •	79.56 ± 0.18
MIFS	66.43 ± 2.27	3	62.30 ± 2.23	83.76 ± 0.19	5 •	79.46 ± 0.21
Naïve Bayes	Full set accuracy	m_{least}	Subset accuracy	Full set accuracy	m_{least}	Subset accuracy
MRmMC	61.67 ± 2.49	3	65.87 ± 2.44	76.13 ± 0.28	2	77.69 ± 0.23
mRMR	60.48 ± 2.61	6 •	57.94 ± 2.66	76.22 ± 0.18	2	76.46 ± 0.15
MIFS	61.59 ± 2.31	7 •	58.17 ± 2.59	76.27 ± 0.21	2	76.32 ± 0.24
SVM	Full set accuracy	m_{least}	Subset accuracy	Full set accuracy	m_{least}	Subset accuracy
MRmMC	63.17 ± 1.98	3	61.27 ± 2.46	79.16 ± 0.22	2	78.34 ± 0.20
mRMR	63.65 ± 2.35	3	65.87 ± 1.59	78.98 ± 0.14	3 •	74.40 ± 0.24
MIFS	64.21 ± 2.03	8 •	62.78 ± 2.53	79.06 ± 0.22	3 •	74.36 ± 0.24
CART	Full set accuracy	m_{least}	Subset accuracy	Full set accuracy	m_{least}	Subset accuracy
MRmMC	66.35 ± 2.52	3	63.10 ± 2.36	81.84 ± 0.22	4	77.41 ± 0.29
mRMR	66.35 ± 2.30	3	64.84 ± 2.22	81.64 ± 0.21	6 •	77.84 ± 0.22
MIFS	68.73 ± 2.41	5 •	66.27 ± 2.45	81.95 ± 0.32	7 •	78.41 ± 0.29
	Vowel			Statlog		
5-NN	Full set accuracy	m_{least}	Subset accuracy	Full set accuracy	m_{least}	Subset accuracy
MRmMC	91.55 ± 0.64	6	87.12 ± 0.82	71.78 ± 0.95	6	67.34 ± 1.04
mRMR	91.73 ± 0.92	6	89.09 ± 0.75	72.13 ± 0.97	9 •	68.93 ± 1.19
MIFS	91.45 ± 0.89	6	87.29 ± 1.00	71.87 ± 1.23	11 •	69.90 ± 1.12
Naïve Bayes	Full set accuracy	m_{least}	Subset accuracy	Full set accuracy	m_{least}	Subset accuracy
MRmMC	73.30 ± 1.19	7	72.73 ± 1.01	60.61 ± 1.25	5	59.03 ± 1.35
mRMR	73.33 ± 1.03	6 □	69.87 ± 1.13	61.44 ± 1.24	7 •	60.06 ± 1.32
MIFS	73.13 ± 1.28	7	71.06 ± 1.28	60.34 ± 1.38	6 •	57.04 ± 1.23
SVM	Full set accuracy	m_{least}	Subset accuracy	Full set accuracy	m_{least}	Subset accuracy
MRmMC	77.81 ± 1.12	8	73.23 ± 1.21	79.59 ± 0.92	16	76.11 ± 0.77
mRMR	78.64 ± 1.18	8	75.57 ± 1.08	79.51 ± 0.89	13 □	76.00 ± 1.02
MIFS	78.42 ± 0.83	8	75.00 ± 1.01	79.57 ± 0.93	12 □	77.57 ± 0.97
CART	Full set accuracy	m_{least}	Subset accuracy	Full set accuracy	m_{least}	Subset accuracy
MRmMC	74.07 ± 1.23	5	71.41 ± 1.11	70.75 ± 0.97	7	68.90 ± 1.43
mRMR	74.75 ± 1.36	4 □	70.42 ± 1.11	70.37 ± 1.14	7	65.64 ± 1.31
MIFS	74.58 ± 1.19	4 □	70.37 ± 1.08	69.57 ± 1.08	5 □	65.03 ± 1.19
	Mfeat Zernike			Sonar		
5-NN	Full set accuracy	m_{least}	Subset accuracy	Full set accuracy	m_{least}	Subset accuracy
MRmMC	80.61 ± 0.48	9	77.03 ± 0.52	78.13 ± 1.80	3	76.34 ± 2.17
mRMR	80.60 ± 0.54	9	77.20 ± 0.65	79.43 ± 1.92	8 •	76.26 ± 1.96
MIFS	80.58 ± 0.49	12 •	75.94 ± 0.60	77.89 ± 2.56	3	73.01 ± 1.74
Naïve Bayes	Full set accuracy	m_{least}	Subset accuracy	Full set accuracy	m_{least}	Subset accuracy
MRmMC	72.33 ± 0.70	6	67.58 ± 0.51	75.61 ± 2.59	2	72.52 ± 2.55
mRMR	72.43 ± 0.68	8 •	70.25 ± 0.72	75.12 ± 2.42	2	71.79 ± 2.25
MIFS	72.58 ± 0.70	8 •	68.69 ± 0.54	76.67 ± 1.41	3 •	75.69 ± 2.66
SVM	Full set accuracy	m_{least}	Subset accuracy	Full set accuracy	m_{least}	Subset accuracy
MRmMC	83.01 ± 0.57	14	78.17 ± 0.72	79.76 ± 2.25	3	78.70 ± 2.59
mRMR	82.53 ± 0.41	9 □	77.64 ± 0.52	76.18 ± 2.47	2 □	72.36 ± 2.36
MIFS	82.47 ± 0.45	15 •	78.38 ± 0.66	77.48 ± 1.86	4 •	72.93 ± 1.87
CART	Full set accuracy	m_{least}	Subset accuracy	Full set accuracy	m_{least}	Subset accuracy
MRmMC	66.58 ± 0.82	8	63.19 ± 0.67	73.01 ± 1.79	3	70.16 ± 2.82
mRMR	66.09 ± 0.64	8	63.74 ± 0.80	72.28 ± 2.30	3	67.40 ± 2.90
MIFS	66.68 ± 0.85	8	62.20 ± 0.81	73.66 ± 2.25	3	69.76 ± 3.06

Tables 2 and 3 summarize the mean of the average classification accuracies based on a number of first selected features. The results presented in rows with $m = 5$, 10, 15, and 30 provide the average classification accuracies of the selected features from 2 to $n_f = \min(m, M)$, respectively, where M is the number of original features. As suggested in [44], the four ranges of the number of selected features in our study here are representative as these choices cover the approximate transitory period where the classification accuracy becomes stable for most of the datasets (see Figs. 2–9). A one-tailed two-sample z-test was conducted for each case of the m values in order to evaluate the null hypothesis (H_0) that "the mean accuracy of the proposed method is greater than the mean accuracy of the compared method" . The recorded p-value is the probability corresponding to the z-test. A significant difference is obtained to support the hypothesis if p is lower than 0.05 (5% significance level). Meanwhile, if p is greater than 0.95 then it can be concluded that the compared method outperforms the proposed method. For ease of viewing, results in the p-value columns are marked with the symbol "$*$" and "■" to indicate that the MRmMC method is statically superior or inferior to the compared method, respectively. The p-value columns which are not highlighted by any symbol indicate that the two methods are comparable.

From Tables 2 and 3, it can be observed that the MRmMC method generally provides either better or comparable classification accuracy in comparison with the other two methods for all classifiers when fewer features (e.g. 2–15 features) are used to represent all the candidate features, except in Vowel and Mfeat Factors. The performance of MRmMC is not as good as mRMR for the Vowel dataset with Nearest Neighbour, Naïve Bayes and SVM clas-

Table 5
The least number of selected features, m_{least}, by MRmMC, mRMR and MIFS methods that gives classification accuracy close to (at most 5% less than the full set accuracy) or better than the full feature set. The symbol "•" (or "□") denotes the proposed method has lower (or larger) value of m_{least} than the compared method. Results are based on Musk and Mfeat Factors datasets.

	Musk			Mfeat Factors		
5-NN	Full set accuracy	m_{least}	Subset accuracy	Full set accuracy	m_{least}	Subset accuracy
MRmMC	88.49 ± 0.96	21	83.89 ± 0.91	96.47 ± 0.26	8	92.20 ± 0.50
mRMR	88.21 ± 1.21	23 •	83.54 ± 1.23	96.55 ± 0.24	7 □	92.34 ± 0.37
MIFS	87.37 ± 1.14	30 •	84.00 ± 1.41	96.63 ± 0.30	9 •	92.17 ± 0.51
Naïve Bayes	Full set accuracy	m_{least}	Subset accuracy	Full set accuracy	m_{least}	Subset accuracy
MRmMC	82.81 ± 1.63	20	77.88 ± 2.37	93.87 ± 0.39	8	89.34 ± 0.64
mRMR	82.14 ± 1.08	17 □	78.76 ± 2.19	94.08 ± 0.39	9 •	89.59 ± 0.38
MIFS	80.91 ± 1.50	20	76.86 ± 1.59	93.87 ± 0.32	10 •	90.03 ± 0.47
SVM	Full set accuracy	m_{least}	Subset accuracy	Full set accuracy	m_{least}	Subset accuracy
MRmMC	85.68 ± 0.99	40	81.47 ± 1.22	97.46 ± 0.25	10	92.79 ± 0.51
mRMR	85.05 ± 1.67	40	80.28 ± 1.61	97.62 ± 0.28	9 □	92.97 ± 0.48
MIFS	85.05 ± 1.27	30 □	80.88 ± 1.20	97.74 ± 0.27	10	93.68 ± 0.50
CART	Full set accuracy	m_{least}	Subset accuracy	Full set accuracy	m_{least}	Subset accuracy
MRmMC	77.09 ± 1.63	5	72.67 ± 1.17	88.38 ± 0.55	9	84.17 ± 0.73
mRMR	78.74 ± 1.76	9 •	75.02 ± 1.37	88.01 ± 0.57	7 □	83.67 ± 0.67
MIFS	77.30 ± 1.97	7 •	75.12 ± 1.69	87.88 ± 0.58	9	83.09 ± 0.59

Table 6
A comparison of win/tie/loss counts of the MRmMC method against the other methods. The counts are based on the results presented in Tables 4 and 5.

Win/tie/lose	mRMR	MIFS
5-NN	4/3/1	5/3/0
Naïve Bayes	4/2/2	5/3/0
SVM	1/3/4	4/2/2
CART	2/4/2	3/3/2
Average	2.75/3/2.25	4.25 2.75/1

sifiers but is comparable to mRMR with CART classifier. Furthermore, MRmMC is only slightly inferior to the MIFS method for the Vowel dataset with Nearest Neighbour classifier.

Considering each classifier used, the MRmMC method is only inferior to either mRMR or MIFS for the Mfeat Factors dataset. Specifically, the MRmMC method shows slightly lower performance than the MIFS method with Naive Bayes classifier yet comparable/better performance with the other three classifiers, while conversely, MRmMC produces comparable performance with the mRMR with Naive Bayes classifier but slightly lower performance with the other three classifiers.

Tables 4 and 5 present the performance of MRmMC, mRMR and MIFS methods, generated by using the least number of selected features, m_{least}, with which a classification accuracy more than or close to that obtain by using the complete dataset (with no more than 5% difference). Results from Tables 4 and 5 are further summarized in Table 6 with an intention to specifically demonstrate the capability of the MRmMC method in representing the full feature set. The win/tie/loss scores reported in Table 6 represent the number of benchmark datasets for which the MRmMC method gives lower/equal/higher number of selected features in comparison to other methods.

As can be seen from Table 6, the MRmMC method performs better than the MIFS for all four classifiers. It performs better for two out of four classifiers and shows comparable performance for the fourth classifier (CART) when compared to the mRMR method but does not perform well with SVM classifier. It can also be noticed that MRmMC gives outstanding performance with Nearest Neighbour and Naïve Bayes classifiers. Based on the average results given in the last row of Table 6, it can be concluded that the MRmMC method is the winner in overall when only a small number of features are required to represent the full feature set.

8. Conclusions

The MRmMC method uses a hill-climbing search structure with a straightforward measurement criterion that makes it simple and easy to implement. It is a filter feature selection method as it uses no specific classification scheme in the selection process, and therefore it works well with popular classifiers such as k-NN, naïve Bayes, SVM and CART.

Although the method may not always find the optimal subset as the search is non-exhaustive, it is shown from the experimental and numerical case studies that the method is competent for feature selection and dimensionality reduction.

As mentioned in Section 5, MRmMC possesses several attractive properties, one of which is that there is no need to pre-specify control parameters as required in MIFS methods, and another important one is that it is relatively easier to implement.

The conditional correlation coefficient defined by (16) can well reveal linear relation between two variables X and Y. It also can reveal nonlinear relation if there is a clear functional relationship between X and Y in the strict sense of word. Therefore, the proposed method can well capture linear relations between features, and can also identify nonlinear relations if features are related to each other in some nonlinear manners. A limitation of MRmMC is that the proposed redundancy measure can is reliable for quantitative features, but cannot effectively evaluate the redundancy between a quantitative and a nominal random variable.

In future work, it is of interest to make use of other measures to assess feature redundancy and combine this idea with the feature relevancy measure applied in this paper. The combination is expected to form a new criterion that can be used to effectively deal with both nominal and quantitative features. It would be also interesting to explore the new criterion with other feature search strategies such as floating search selection and nature-inspired selection in order to find better feature subset solutions.

Acknowledgements

Azlyna Senawi gratefully acknowledges a scholarship from the Ministry of Higher Education Malaysia. The authors gratefully acknowledge that part of this work was supported by the Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/I011056/1 and Platform Grant EP/H00453X/1, and EU Horizon 2020 Research and Innovation Action Framework Programme under Grant No 637302 (PROGRESS).

References

[1] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemom. Intell. Lab. Syst. 2 (1) (1987) 37–52.
[2] S. Balakrishnama, A. Ganapathiraju, Linear Discriminant Analysis-A Brief Tutorial, Institute for Signal and information Processing, 1998.
[3] P. Somol, J. Novovicová, P. Pudil, Efficient feature subset selection and subset size optimization, Pattern Recognition Recent Advances, InTech, 2010.
[4] H.-L. Wei, S. Billings, Feature subset selection and ranking for data dimensionality reduction, Pattern Anal. Mach. Intell. IEEE Trans. 29 (1) (2007) 162–166.
[5] M. Dash, H. Liu, Feature selection for classification, Intell. Data Anal. 1 (1) (1997) 131–156.
[6] Y. Peng, Z. Wu, J. Jiang, A novel feature selection approach for biomedical data classification, J. Biomed. Inf. 43 (1) (2010) 15–23.
[7] P. Pudil, J. Novovičová, J. Kittler, Floating search methods in feature selection, Pattern Recognit. Lett. 15 (11) (1994) 1119–1125.
[8] P. Somol, P. Pudil, J. Novovičová, P. Paclık, Adaptive floating search methods in feature selection, Pattern Recognit. Lett. 20 (11) (1999) 1157–1163.
[9] F.M. Lopes, D.C. Martins, J. Barrera, R.M. Cesar, A feature selection technique for inference of graphs from their known topological properties: revealing scale-free gene regulatory networks, Inf. Sci. 272 (2014) 1–15.
[10] D.B. Skalak, Prototype and feature selection by sampling and random mutation hill climbing algorithms, in: Proceedings of the Eleventh International Conference on Machine Learning, 1994, pp. 293–301.
[11] W. Siedlecki, J. Sklansky, A note on genetic algorithms for large-scale feature selection, Pattern Recognit. Lett. 10 (5) (1989) 335–347.
[12] J. Yang, V. Honavar, Feature subset selection using a genetic algorithm, in: Feature Extraction, Construction and Selection, Springer, 1998, pp. 117–136.
[13] S. Tabakhi, P. Moradi, Relevance–redundancy feature selection based on ant colony optimization, Pattern Recognit. 48 (9) (2015) 2798–2811.
[14] S.-W. Lin, K.-C. Ying, S.-C. Chen, Z.-J. Lee, Particle swarm optimization for parameter determination and feature selection of support vector machines, Expert Syst. Appl. 35 (4) (2008) 1817–1824.
[15] M.A. Hall, Correlation-based Feature Selection For Machine Learning, The University of Waikato, 1999.
[16] G.H. John, R. Kohavi, K. Pfleger, Irrelevant features and the subset selection problem, in: Machine Learning: Proceedings of the Eleventh International Conference, 1994, pp. 121–129.
[17] D. Koller and M. Sahami, "Toward optimal feature selection," 1996.
[18] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (Dec (1-2)) (1997) 273–324.
[19] L. Yu, H. Liu, Efficient feature selection via analysis of relevance and redundancy, J. Mach. Learn. Res. 5 (2004) 1205–1224.
[20] R. Battiti, Using mutual information for selecting features in supervised neural net learning, Neural Netw. IEEE Trans. 5 (4) (1994) 537–550.
[21] N. Kwak, C.-H. Choi, Input feature selection for classification problems, Neural Netw. IEEE Trans. 13 (1) (2002) 143–159.
[22] P. Estévez, M. Tesmer, C. Perez, J.M. Zurada, Normalized mutual information feature selection, Neural Netw. IEEE Trans. 20 (2) (2009) 189–201.
[23] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, Pattern Anal. Mach. Intell. IEEE Trans. 27 (8) (2005) 1226–1238.
[24] W. Li, Mutual information functions versus correlation functions, J. Stat. Phys. 60 (5-6) (1990) 823–837.
[25] R. Moddemeijer, On estimation of entropy and mutual information of continuous distributions, Signal Process. 16 (3) (1989) 233–248.
[26] Y.-I. Moon, B. Rajagopalan, U. Lall, Estimation of mutual information using kernel density estimators, Phys. Rev. E 52 (3) (1995) 2318–2321.
[27] A. Kraskov, H. Stögbauer, P. Grassberger, Estimating mutual information, Phys. Rev. E 69 (6) (2004) 066138.
[28] N. Kwak, C.-H. Choi, Input feature selection by mutual information based on Parzen window, Pattern Anal. Mach. Intell. IEEE Trans. 24 (12) (2002) 1667–1671.
[29] C.O. Daub, R. Steuer, J. Selbig, S. Kloska, Estimating mutual information using B-spline functions–an improved similarity measure for analysing gene expression data, BMC Bioinf. 5 (1) (2004) 118.
[30] A.M. Fraser, H.L. Swinney, Independent coordinates for strange attractors from mutual information, Phys. Rev. A 33 (2) (1986) 1134.
[31] G.A. Darbellay, I. Vajda, Estimation of the information by an adaptive partitioning of the observation space, IEEE Trans. Inf. Theory 45 (4) (1999) 1315–1321.
[32] D. Yu, S. An, Q. Hu, Fuzzy mutual information based min-redundancy and max-relevance heterogeneous feature selection, Int. J. Comput. Intell. Syst. 4 (4) (2011) 619–633.
[33] J. Walters-Williams, Y. Li, Estimation of mutual information: A survey, in: Rough Sets and Knowledge Technology, Springer, 2009, pp. 389–396.
[34] C. Ding, H. Peng, Minimum redundancy feature selection from microarray gene expression data, J. Bioinf. Comput. Biol. 3 (02) (2005) 185–205.
[35] Y. Wang, B. Tan, Y. Wang, J. Wu, Information structure analysis for quantitative assessment of mineral resources and the discovery of a silver deposit, Nonrenewable Resour. 3 (4) (1994) 284–294.
[36] D.C. Whitley, M.G. Ford, D.J. Livingstone, Unsupervised forward selection: a method for eliminating redundant variables, J. Chem. Inf. Comput. Sci. 40 (5) (2000) 1160–1168.
[37] R.F. Tate, Correlation between a discrete and a continuous variable: point-biserial correlation, Ann. Math. Stat. 25 (3) (1954) 603–607.
[38] S.-Y. Jiang, L.-X. Wang, Efficient feature selection based on correlation measure between continuous and discrete features, Inf. Process. Lett. (2015).
[39] Q. Gu, Z. Li, and J. Han, "Generalized fisher score for feature selection," arXiv preprint arXiv:1202.3725, 2012.
[40] G.A. Abandah, T.M. Malas, Feature selection for recognizing handwritten Arabic letters, Dirasat Eng. Sci. J. 37 (2) (2010).
[41] S.A. Billings, Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains, John Wiley & Sons, 2013.
[42] 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 (1) (2012) 27–66.
[43] X. Wu, et al., Top 10 algorithms in data mining, Knowl. Inf. Syst. 14 (1) (2008) 1–37.
[44] J.M. Sotoca, F. Pla, Supervised feature selection by clustering using conditional mutual information-based distances, Pattern Recognit. 43 (6) (2010) 2068–2081.

Azlyna Senawi received her B.Sc. degree in Industrial Mathematics and M.Sc. in Mathematics, both from Universiti Teknologi Malaysia in 2003 and 2005, respectively. She is currently a Ph.D. student in the Department of Automatic Control and Systems Engineering, the University of Sheffield. Her research interests include data mining and pattern recognition.

Hua-Liang Wei (B.Sc., M.Sc., Ph.D.) is a senior lecturer in the Department of Automatic Control and Systems Engineering, University of Sheffield. His recent research interests include system identification and data analytics (SIDA).

Stephen A. Billings is a professor in the Department of Automatic Control and Systems Engineering, University of Sheffield, UK, and leads the Signal Processing and Complex Systems research group. His research interests include system identification, nonlinear system modelling, design and analysis, with applications in many multidisciplinary areas.