SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

January 2020, Vol. 63 112103:1–112103:15
https://doi.org/10.1007/s11432-019-2633-y

# On some aspects of minimum redundancy maximum relevance feature selection

Peter BUGATA & Peter DROTAR*

*Intelligent Information Systems Lab, Technical University of Kosice, Kosice* 04013*, Slovakia*

Received 25 April 2019/Revised 12 July 2019/Accepted 21 August 2019/Published online 24 December 2019

**Abstract** The feature selection is an important challenge in many areas of machine learning because it plays a crucial role in the interpretations of machine-driven decisions. There are various approaches to the feature selection problem and methods based on the information theory comprise an important group. Here, the minimum redundancy maximum relevance (mRMR) feature selection is undoubtedly the most popular one with widespread application. In this paper, we prove in contrast to an existing finding that the mRMR is not equivalent to Max-Dependency criterion for first-order incremental feature selection. We present another form of equivalence leading to a generalization of mRMR feature selection. Additionally, we compare several feature selection methods based on mRMR, Max-Dependency, and feature ranking, employing different measures of dependency. The results on high-dimensional real-world datasets show that the distance correlation is the suitable measure for dependency-based feature selection methods. The results also indicate that the Max-Dependency incremental algorithm combined with distance correlation appears to be a promising feature selection approach.

**Keywords** big data, information theory, feature selection, dimensionality reduction, minimum redundancy maximum relevance, mRMR

**Citation** Bugata P, Drotar P. On some aspects of minimum redundancy maximum relevance feature selection. Sci China Inf Sci, 2020, 63(1): 112103, https://doi.org/10.1007/s11432-019-2633-y

## 1 Introduction

In recent years, there has been an unprecedented growth in data size. The number of samples in datasets is growing because the data acquisition in many areas such as social networks and multimedia is more feasible than ever before. Not only sample size, but also dimensionality is increasing rapidly. There are areas of research such as bioinformatics and image processing where the dimensionality of tens or hundreds of thousands of features is quite frequent. Although the development of algorithms is following the trend determined by the data growth and new algorithms are being developed, there are still several challenges with big and high-dimensional data. Some examples of a challenging scenario are high-dimensional and small sample datasets, where the dimensionality of the data is significantly higher than the sample size [1, 2]. In predictive modeling, high-dimensionality can lead to overfitting when the model captures the patterns in the training data well but fails to generalize well to unseen data. In a setup like this, it is necessary to reduce the dimensionality of the data to avoid effects of the curse of dimensionality. These include overfitting, prolonged computational times, and negative influence on classification algorithms.

There are two ways to dimensionality reduction: feature selection and feature transformation [3]. Feature selection (FS) is the process of obtaining a subset from the original feature set, which selects the

* Corresponding author (email: peter.drotar@tuke.sk)

© Science China Press and Springer-Verlag GmbH Germany, part of Springer Nature 2019 

relevant features of the dataset and removes the irrelevant and redundant ones. In contrast, feature transformation achieves dimensionality reduction by combining the original features into a set of new features with stronger discriminating power. Each of the approaches has its advantages. Feature transformation is preferable in applications where model accuracy is more important than model interpreting. Because FS gives the original features, it is especially useful for model interpreting and knowledge extraction.

The FS methods can be categorized based on the availability of supervision to supervised, unsupervised, and semi-supervised methods. Concerning different FS strategies, we distinguish between filter, wrapper, and embedded methods. Finally, from the data perspective, the FS can be categorized as FS on static data and FS on streaming data. The description of different approaches is beyond the scope of this paper, but there are excellent survey papers such as [1, 4, 5] that provided additional information.

We focus on the most popular supervised FS. Supervised FS is the process of selecting a feature subset based on some criteria for measuring the importance and relevance of the features by utilizing the target variable to train the FS model [4]. It is usually taken as a pre-processing step for the classification or regression tasks. It chooses features that can distinguish data samples from different classes or samples with different regression target values [2].

Filter methods are independent of any learning algorithm, therefore, they can provide general solutions usable for various learning methods. They are advantageous for their low computational cost and simplicity. Filters rely on the intrinsic characteristics of data based on dependency, information, distance, and consistency. Recently, some methods have been proposed to select features based on mutual information. In information theory, the mutual information of two random variables determines the amount of information obtained about one random variable by observing the other random variable. It can be used to measure the dependency between two variables which in general can be multidimensional. Owing to the difficulty in computing the mutual information of the multidimensional variables, some FS methods estimate it by using the linear combinations of the information terms for two one-dimensional variables, for example, in [6, 7], or by incorporating the conditional and joint mutual information [8–11]. A review of the state-of-the-art of information-theoretic FS methods can be found in [12, 13].

Probably the most popular[1)] from all available FS methods that have widespread application is minimum redundancy maximum relevance (mRMR) FS [15]. It is widely used in many domains such as bioinformatics and multimedia processing [16, 17]. The mRMR FS is a supervised filter method with a forward selection that uses mutual information as a dependency measure. The method iteratively extends the set of selected explanatory variables by maximizing relevance toward the target variable and at the same time, minimizing the redundancy among the selected variables.

In this paper, we focus on several aspects of mRMR FS. In [14], the authors presented Theorem 1 claiming that mRMR is equivalent to maximal dependency (Max-Dependency) for the first-order incremental search. We show that this theorem does not hold. Using a synthetically constructed discrete dataset, we demonstrate the shortcomings of the proof described in [14]. To answer the question of mRMR equivalence, we introduce an objective function whose maximization is equivalent to the mRMR algorithm. This function allows generalizing the mRMR algorithm by applying various weights of the average redundancy of the selected variables.

It is known that the performance of the mRMR method is influenced by the dependency measure used for evaluating the relevance and redundancy of considered variables [18]. Therefore, besides the mutual information as an original measure, we incorporate other dependency measures into our numerical experiments. These new mRMR versions are compared to other dependency-based FS methods, including the maximal dependency incremental algorithm.

The rest of the paper is organized as follows: Section 2 presents a summary of the mRMR algorithm. Section 3 focuses on theoretical properties of mRMR. We present a counterexample to the theorem about the equivalence of mRMR and Max-Dependency and define the objective function, maximization of which is equivalent to mRMR. Section 4 describes dependency measures used in the numerical experiments and analyzes the obtained results. Finally, Section 5 presents our conclusion.

1) There are more than 3400 Web of Science citations to the Peng's paper [14].

## 2 Minimum redundancy maximum relevance

The mRMR FS method is a filter with a forward selection. In every step, the method extends the set of selected explanatory variables (features) by the next one to maximize relevance toward the target variable and simultaneously to minimize the redundancy among the already selected variables. The original method is proposed in [14, 15].

Let dataset be represented by a two-dimensional matrix, whose rows are the observations and the columns are the explanatory variables. We denote $\{x_1, x_2, \ldots, x_n\}$ as the set of $n$ observations and $\{X_1, X_2, \ldots, X_k\}$ as the set of $k$ explanatory variables or features. Let the target variable $Y$ be a vector $(y_1, y_2, \ldots, y_n)$. Consider the general case when the values of explanatory variables are real numbers, then the variables constitute the $k$-dimensional space $\mathbb{R}^k$. The observations are data points in this high-dimensional space.

The FS aims to find the subset $S$ of explanatory variables that optimally characterize the target variable $Y$. Under optimal characterization, we will understand the highest statistical dependency of the target variable $Y$ on the selected subset $S$. This criterion is known as a maximal dependency.

In mRMR [15], the mutual information is utilized as a dependency measure of two random variables. The mutual information between the set of explanatory variables $S$ and the target variable $Y$ is defined as

$$I(S;Y) = \iint p(S,Y) \log \frac{p(S,Y)}{p(S)p(Y)} \,\mathrm{d}S \,\mathrm{d}Y, \tag{1}$$

where $p(Y)$ is the probability density function (PDF) of the continuous random variable $Y$. The terms $p(S)$ and $p(S,Y)$ stand for the joint PDFs of the continuous random variable sets $S$ and $S \cup \{Y\}$, respectively.

In the case of continuous random variables, the calculation of the mutual information is difficult because the densities are all unknown and are hard to estimate from data. Therefore, discretizing the values or density approximation is preferred [19]. In the case of the discrete variables $Z$ and $Y$, the mutual information is defined as follows [19]:

$$I(Z;Y) = \sum_{z_i} \sum_{y_j} P(Z = z_i, Y = y_j) \log \frac{P(Z = z_i, Y = y_j)}{P(Z = z_i)P(Y = y_j)}. \tag{2}$$

The probabilities $P$ are determined from the contingency tables covering count of values occurrences of the variables $Z$ and $Y$.

The goal of the FS algorithm is to find the subset $S$ of the explanatory variables on which $Y$ shows the highest dependency expressed by [15] using mutual information as

$$\max D(S,Y), \quad D = I(S;Y). \tag{3}$$

Because the Max-Dependency criterion is hard to implement, the preferred alternative is to use maximum relevance (Max-Relevance) as a criterion. The Max-Relevance condition looks for the highest dependency of the target variable on the individual selected variables, i.e., it searches for the variables satisfying

$$\max D(S,Y), \quad D = \frac{1}{|S|} \sum_{X_i \in S} I(X_i;Y). \tag{4}$$

Eq. (4) approximates Max-Dependency by the highest average dependency of the target variable $Y$ on the individual variables from the set $S$. However, the explanatory variables selected by this approach can be redundant. If one of the redundant variables is omitted, the prediction performance does not decrease. Therefore, the condition for minimum redundancy (Min-Redundancy) to drop redundant variables is added:

$$\min R(S), \quad R = \frac{1}{|S|^2} \sum_{X_i, X_j \in S} I(X_i;X_j). \tag{5}$$

The criterion that combines both conditions is the mRMR. The following function represents the simplest way to concurrently optimize relevance $D$ and redundancy $R$:

$$\max \Phi(D, R), \quad \Phi = D - R, \quad \text{respectively} \quad \Phi = \frac{D}{R}. \tag{6}$$

Sequential forward search (first-order incremental search) can be used to find the subset of the variables close to the optimal variables obtained by maximizing the function $\Phi$ defined by (6). In the first step, the variable with the largest relevance toward the target variable is selected. When the subset $S$ with $m$ variables is known, then the $(m+1)$th variable is selected from the other variables by optimizing one of the following conditions [15]:

- For subtraction (MID—mutual information difference criterion):

$$\max_{X_i \notin S} \left[ I(X_i; Y) - \frac{1}{m} \sum_{X_j \in S} I(X_i; X_j) \right]; \tag{7}$$

- For division (MIQ—mutual information quotient criterion):

$$\max_{X_i \notin S} \left[ I(X_i; Y) \Big/ \frac{1}{m} \sum_{X_j \in S} I(X_i; X_j) \right]. \tag{8}$$

This step is repeated until the required number of variables is selected. The order of selection of the variables corresponds to their ranking by importance.

## 3 Theoretical aspects of mRMR feature selection

In this section, we focus on the theoretical properties of the mRMR algorithm. First, we analyze the equivalence between mRMR and Max-Dependency presented in [14]. We construct a dataset on which these two FS methods give different results using the first-order incremental search. Similar findings are presented on some real-world datasets. Then, we examine the relationship between the objective function $\Phi$ defined in (6) and the incremental step of the mRMR FS method (7). We define a new objective function of feature subsets, maximization of which is equivalent to the mRMR algorithm for the first-order incremental search. The definition of the objective function is the basis for generalization of the mRMR FS method.

### 3.1 A counterexample to the equivalence of mRMR and Max-Dependency

As was already indicated, the algorithm for Max-Dependency is computationally demanding especially in the case of continuous variables. The mRMR algorithm represents a computationally more feasible approach. Peng et al. [14] claims the equivalence of mRMR and Max-Dependency.

**Theorem 1.** For the first-order incremental search, mRMR is equivalent to Max-Dependency.

The motivation for a counterexample to the theorem comes from [19] showing an example of two independent variables $X_1, X_2$ such that $I(X_1; Y) = 0$ and $I(X_2; Y) = 0$, whereas the combination of these two allows to easily differentiate between different classes of the target variable.

Table 1 provides description of the binary classification dataset containing 16 observations. All three explanatory variables are discrete. The variable $X_1$ takes three values, whereas $X_2, X_3$ take two values. We will examine how the mRMR and Max-Dependency methods work in selecting a two-element set of variables.

It can be shown that for this dataset, the following holds:

$$I(X_1; X_3) = 0, \quad I(X_2; X_3) = 0.$$

**Table 1** Dataset to compare mRMR and Max-Dependency algorithms

| | $X_1$ | $X_2$ | $X_3$ | Y | | $X_1$ | $X_2$ | $X_3$ | Y |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 9 | 0 | 0 | 1 | 0 |
| 2 | 0 | 0 | 0 | 0 | 10 | 0 | 0 | 1 | 0 |
| 3 | 0 | 1 | 0 | 1 | 11 | 0 | 1 | 1 | 1 |
| 4 | 0 | 1 | 0 | 1 | 12 | 0 | 1 | 1 | 1 |
| 5 | 1 | 0 | 0 | 1 | 13 | 1 | 0 | 1 | 1 |
| 6 | 1 | 0 | 0 | 1 | 14 | 1 | 0 | 1 | 1 |
| 7 | 1 | 1 | 0 | 0 | 15 | 1 | 1 | 1 | 0 |
| 8 | 2 | 1 | 0 | 0 | 16 | 2 | 1 | 1 | 0 |

**Table 2** Contingency tables for counterexample to Theorem 1

| $(X_1, X_2)$ | $Y=0$ | $Y=1$ | $\Sigma$ | $(X_1, X_3)$ | $Y=0$ | $Y=1$ | $\Sigma$ |
|---|---|---|---|---|---|---|---|
| (0, 0) | 4 | 0 | 4 | (0, 0) | 2 | 2 | 4 |
| (1, 0) | 0 | 4 | 4 | (1, 0) | 1 | 2 | 3 |
| (2, 0) | 0 | 0 | 0 | (2, 0) | 1 | 0 | 1 |
| (0, 1) | 0 | 4 | 4 | (0, 1) | 2 | 2 | 4 |
| (1, 1) | 2 | 0 | 2 | (1, 1) | 1 | 2 | 3 |
| (2, 1) | 2 | 0 | 2 | (2, 1) | 1 | 0 | 1 |
| $\Sigma$ | 8 | 8 | 16 | $\Sigma$ | 8 | 8 | 16 |

Additionally,

$$I(X_2;Y)=0, \quad I(X_3;Y)=0.$$

The mRMR and Max-Dependency algorithms select variable $X_1$ in the first step because the value of $I(X_1;Y)$ is positive (approximately 0.156).

In the second step, mRMR calculates the differences of $I()$ by (7) for the candidates:

$$X_2 : I(X_2;Y) - I(X_2;X_1) = -I(X_2;X_1),$$

$$X_3 : I(X_3;Y) - I(X_3;X_1) = 0 - 0 = 0.$$

Because the value of $I(X_2;X_1) \cong 0.156$ is positive, the maximum is achieved for $X_3$ and mRMR chose $X_3$ as a second selected variable. However, this choice is not correct because $X_3$ is not relevant for prediction of the target variable.

On the other hand, Max-Dependency selects $X_2$ as a second selected variable. The selection by Max-Dependency is based on the mutual information values of the sets $\{X_1, X_2\}$ and $\{X_1, X_3\}$ with the target variable $Y$.

The resulting values for the candidates are

$$X_2 : I(\{X_1, X_2\};Y) = 1,$$

$$X_3 : I(\{X_1, X_3\};Y) = I(X_1;Y) \cong 0.156.$$

So in this case $I(\{X_1, X_2\};Y) > I(\{X_1, X_3\};Y)$.

Therefore, for this particular dataset, the Max-Dependency algorithm selects the different pair of variables than the mRMR algorithm when the first-order incremental search is used. In addition, the pair selected by the Max-Dependency algorithm shows substantially larger dependency on the target variable as the pair selected by mRMR. This is the counterexample to Peng's theorem stated in [14].

For an explanation, we describe how we calculate the mutual information values of the sets $\{X_1, X_2\}$ and $\{X_1, X_3\}$ with the target $Y$. The mutual information is determined according to (2) by substituting the pairs $(X_1, X_2)$ and $(X_1, X_3)$ for $Z$ using Table 2.

Now, we outline the main idea of the original proof of Theorem 1. Based on the definition of the first-order search, the authors assume that the set $S_{m-1}$ of $m-1$ variables has already been obtained

and the task is to select the optimal $m$th variable $X_m$. The Max-Dependency algorithm searches the variable $X_m \notin S_{m-1}$ to maximize $D = I(S_m; Y)$, where $S_m = S_{m-1} \cup \{X_m\}$.

The authors show that the mutual information $I(S_m; Y)$ can be expressed as a difference of two terms:

$$I(S_m; Y) = J(S_m, Y) - J(S_m), \tag{9}$$

where $J(S_m)$ for $S_m = \{X_1, \ldots, X_m\}$ is a "mutual information" for multiple scalar variables defined as

$$J(S_m) = \int \cdots \int p(X_1, \ldots, X_m) \log \frac{p(X_1, \ldots, X_m)}{p(X_1) \cdots p(X_m)} \, \mathrm{d}X_1 \ldots \mathrm{d}X_m. \tag{10}$$

The authors claim that Max-Dependency expressed by maximizing the left side of equation (9) is equivalent to simultaneously maximizing the first term and minimizing the second term of the subtraction on the right side of equation (9). They derive that the maximum of the first term on the right side of equation (9) is attained when all the variables from $S_m \cup \{Y\}$ are maximally dependent. Because $S_{m-1}$ is fixed, the $X_m$ and $Y$ should have the maximal dependency. This is the Max-Relevance criterion. The minimum of the second term is attained when all variables of $S_m$ are independent of each other. As all the $m-1$ variables have been selected, this pair-wise independence condition means that the mutual information between $X_m$ and any selected variable from $S_{m-1}$ is minimized. This is the Min-Redundancy criterion.

The authors conclude that mRMR, as a combination of Max-Relevance and Min-Redundancy, is equivalent to Max-Dependency for first-order selection.

However, considering the abovementioned counterexample, when the mRMR and Max-Dependency algorithms select the second variable to build the set $S_2$, the following holds for candidates:

$X_2$: Max-Dependency:

$$I(\{X_1, X_2\}; Y) = J(X_1, X_2, Y) - J(X_1, X_2) \cong 1.156 - 0.156 = 1,$$

whereas mRMR: $I(X_2; Y) - I(X_1; X_2) = 0 - 0.156 = -0.156$;

$X_3$: Max-Dependency:

$$I(\{X_1, X_3\}; Y) = J(X_1, X_3, Y) - J(X_1, X_3) \cong 0.156 - 0 = 0.156,$$

whereas mRMR: $I(X_3; Y) - I(X_1; X_3) = 0 - 0 = 0$.

We observe that the second term in the subtraction for the individual candidates is the same for Max-Dependency and mRMR. But the difference is in the first term. $J(X_1, X_2, Y)$ and $J(X_1, X_3, Y)$ cannot be replaced by the terms $I(X_2; Y)$ and $I(X_3; Y)$, respectively, because they reflect the joint effect of variables on the target class, whereas the terms $I(X_2; Y)$ and $I(X_3; Y)$ show only independence of the target variable $Y$ from the individual variables $X_2$ and $X_3$. This is the reason why the Max-Dependency and mRMR algorithms give different results of variable selection on the constructed dataset.

It is worth noting that the original proof assumes that if $S_{m-1}$ is fixed, then there exists one variable $X_m \notin S_{m-1}$ simultaneously maximizing the first term and minimizing the second term of the subtraction on the right side of (9). But this is not always true. The presented counterexample shows that if $S_1 = \{X_1\}$ is fixed, then in the second step, the first term on the right side of (9) is maximized by $X_2$ because $J(X_1, X_2, Y) = 1.156$ and $J(X_1, X_3, Y) = 0.156$, whereas the second term is minimized by $X_3$ because $J(X_1, X_2) = 0.156$ and $J(X_1, X_3) = 0$.

It can be proven that the similar situation as for the constructed dataset occurs also for real-world datasets. We study the behavior of the Max-Dependency and mRMR FS methods for the first-order incremental search on high-dimensional microarray datasets described in Table 3 [20–27].

Let us consider the Golub dataset [20]. Continuous features are discretized in the similar way as in [15] to three discrete values using thresholds mean − std and mean + std. Here, mean stands for the average value of a feature and std represents for the standard deviation of a particular feature. For this dataset, the Max-Dependency selects four features v3319, v803, v6942, and v48 (in this order) with the mutual information between this feature set and the target variable equal to 0.93. In contrast, mRMR selects

**Table 3** Characteristics of datasets used in this study

| Dataset (Source) | Number of samples | Number of features | Number of Class 0 | Number of Class 1 |
|---|---|---|---|---|
| Alon [22] | 62 | 2000 | 40 | 22 |
| Tian [23] | 173 | 12625 | 36 | 137 |
| Gordon [21] | 181 | 12533 | 94 | 87 |
| Golub [20] | 72 | 7129 | 47 | 25 |
| Burczynski [24] | 127 | 22283 | 85 | 42 |
| Pomeroy [25] | 60 | 7128 | 39 | 21 |
| Singh [26] | 102 | 12600 | 52 | 50 |
| Chowdary [27] | 104 | 22283 | 62 | 42 |

**Table 4** Contingency table for Gordon dataset

| (v12113, v3333, v3249) | $Y=0$ | $Y=1$ | $\sum$ |
|---|---|---|---|
| (0,1,0) | 17 | 0 | 17 |
| (0,1,2) | 4 | 0 | 4 |
| (0,2,0) | 9 | 0 | 9 |
| (0,2,2) | 0 | 7 | 7 |
| (1,2,0) | 0 | 1 | 1 |
| (1,2,1) | 0 | 5 | 5 |
| (1,2,2) | 0 | 11 | 11 |
| (2,1,2) | 1 | 0 | 1 |
| (2,2,0) | 0 | 1 | 1 |
| (2,2,1) | 0 | 23 | 23 |
| (2,2,2) | 0 | 102 | 102 |
| $\sum$ | 31 | 150 | 181 |

the first four iteration features v3319, v803, v4846, and v1828, i.e., there is a difference in third and also fourth selected features. Moreover, the mutual information of the selected set of features with the target variable is only 0.84 that is considerably less than the mutual information obtained by the set selected by Max-Dependency. Also, the Max-Dependency algorithm selects only four features as an optimal feature set. Adding more features do not increase the mutual information between the set of selected features and the target variable.

The similar behavior is observed for the Gordon dataset [21]. In this case, only three features are selected by Max-Dependency. The contingency table (Table 4) shows that selecting only three out of 12533 features is enough to determine the value of the target variable $Y$.

The described examples show one drawback of mRMR as a bivariate FS method. Bivariate methods are not expected to perform as well as multivariate ones because they only consider pair-wise redundancy or relevance. Thus, this algorithm is not able to select features which together provide more information about the target variable than any one separately.

### 3.2 Generalization of mRMR algorithm

By comparing definitions (7) and (8) of the incremental step of mRMR proposed by [15] and the definition of the objective function $\Phi$ in (6), we recognize several differences. For example, incremental mRMR does not contain the expression $I(X_i;X_j)$ for $i=j$. It can be experimentally shown that the results yielded by mRMR based on the incremental definition are different from the results obtained by the incremental algorithm for maximization of the function $\Phi$.

The following lemma introduces the objective function $\Phi'$ of variable subsets. We prove by induction that the first-order incremental algorithm for maximization of $\Phi'$ selects the same set of variables as mRMR in mutual information difference criterion (MID) version with the incremental definition (7).

**Lemma 1.** Let the function $\Phi'$ be defined as follows:

$$\Phi'(S) = \frac{1}{|S|} \sum_{X_i \in S} I(X_i; Y) - \frac{1}{2} \cdot \frac{1}{|S|\,(|S|-1)} \sum_{\substack{X_i, X_j \in S, \\ i \neq j}} I(X_i; X_j). \tag{11}$$

For the first-order incremental search, mRMR algorithm in MID implementation is equivalent to the algorithm for maximization of function $\Phi'$.

*Proof.* In the first step, the incremental implementation of the mRMR algorithm selects the variable with maximal relevance to the target variable. The same is true for the incremental algorithm for maximization of the function $\Phi'$. We will show that if both algorithms in the first $m$ steps select the same variables into the subset $S$, then they select the same variable $X$ in step $m+1$.

Let $S' = S \cup \{X\}$. The set $S'$ has $m+1$ items. The value of the optimized function after addition of $X$ is

$$\Phi'(S') = \frac{1}{m+1} \sum_{X_i \in S'} I(X_i; Y) - \frac{1}{2(m+1)m} \sum_{\substack{X_i, X_j \in S', \\ i \neq j}} I(X_i; X_j). \tag{12}$$

The algorithm for maximization of the function $\Phi'$ selects variable $X$ to maximize the expression on the right side of (12), whose terms can be expressed as follows:

$$\frac{1}{m+1} \sum_{X_i \in S'} I(X_i; Y) = \frac{1}{m+1} \left( \sum_{X_i \in S} I(X_i; Y) + I(X; Y) \right), \tag{13}$$

$$\frac{1}{2(m+1)m} \sum_{\substack{X_i, X_j \in S', \\ i \neq j}} I(X_i; X_j) = \frac{1}{2(m+1)m} \left( \sum_{\substack{X_i, X_j \in S, \\ i \neq j}} I(X_i; X_j) + 2 \sum_{X_i \in S} I(X_i; X) \right). \tag{14}$$

After substitution, we obtain

$$\begin{aligned} \Phi'(S') = {} & \frac{1}{m+1} \sum_{X_i \in S} I(X_i; Y) + \frac{1}{m+1} I(X; Y) \\ & - \frac{1}{2(m+1)m} \sum_{\substack{X_i, X_j \in S, \\ i \neq j}} I(X_i; X_j) - \frac{1}{(m+1)m} \sum_{X_i \in S} I(X_i; X). \end{aligned} \tag{15}$$

The terms in (15) not containing the variable $X$ do not have any influence on maximization of $\Phi'(S')$ with respect to $X$. Therefore, it is enough to maximize the expression:

$$\frac{1}{m+1} I(X; Y) - \frac{1}{(m+1)m} \sum_{X_i \in S} I(X_i; X). \tag{16}$$

The positive term $\frac{1}{m+1}$ may be taken out and the resulting expression for maximization is as follows:

$$\max_{X \notin S} \left[ I(X; Y) - \frac{1}{m} \sum_{X_i \in S} I(X_i; X) \right]. \tag{17}$$

The expression (17) is equivalent to the incremental step (7) in MID mRMR. Therefore, the first-order incremental search for maximization of the function $\Phi'$ selects the same variable $X$ as the incremental mRMR algorithm.

The function $\Phi'$ is the difference between the average dependency of the explanatory variables on the target variable and the average dependency between different variables in the set $S$ to each other weighted by $\frac{1}{2}$. The function can be generalized as follows:

$$\Phi'(S) = \frac{1}{|S|} \sum_{X_i \in S} I(X_i; Y) - \frac{\lambda}{|S| \cdot (|S|-1)} \sum_{\substack{X_i, X_j \in S, \\ i \neq j}} I(X_i; X_j), \tag{18}$$

where $\lambda$ is a positive real number which expresses the weight of redundancy.

It can be proven that the value of $\lambda = \frac{1}{2}$ is not always the best choice. Consider a dataset that is created from the counterexample presented in Table 1 changing the value of the variable $X_2$ in the 16th observation from 1 to 0. When comparing to the original dataset, this change results in decreasing dependency between the variables $X_1$ and $X_2$ and increasing the dependency of $X_3$ and $Y$ on $X_2$.

We show that maximizing the objective function $\Phi'$ with $\lambda = \frac{1}{2}$, i.e., the mRMR algorithm, does not select the optimal two-element subset of variables on this dataset.

It can be shown that for this dataset, the following holds:

$$I(X_1; X_2) = 0.019, \quad I(X_1; X_3) = 0, \quad I(X_2; X_3) = 0.011.$$

Additionally, $I(X_1; Y) = 0.156$, $I(X_2; Y) = 0.011$, $I(X_3; Y) = 0$.

The algorithm for maximization of the function $\Phi'$ with $\lambda = \frac{1}{2}$ selects the variable $X_1$ with the maximal dependency on the target $Y$ in the first step. In the second step, the values of $\Phi'$ for the subsets $\{X_1, X_2\}$ and $\{X_1, X_3\}$ are calculated by (18) as follows:

$$\begin{aligned} \Phi'(\{X_1, X_2\}) &= \frac{1}{2}(I(X_1; Y) + I(X_2; Y)) - \frac{1}{4}(I(X_1; X_2) + I(X_2; X_1)) \cong 0.074, \\ \Phi'(\{X_1, X_3\}) &= \frac{1}{2}(I(X_1; Y) + I(X_3; Y)) - \frac{1}{4}(I(X_1; X_3) + I(X_3; X_1)) \cong 0.078. \end{aligned} \tag{19}$$

The maximum is achieved for $\{X_1, X_3\}$, thus the algorithm for maximizing the function $\Phi'$, like mRMR, selects $X_3$ as a second variable. On the other hand, Max-Dependency selects $X_2$, because the values of the mutual information of the sets $\{X_1, X_2\}$ and $\{X_1, X_3\}$ with the target variable $Y$ are $I(\{X_1, X_2\}; Y) = 1$ and $I(\{X_1, X_3\}; Y) \cong 0.156$.

It implies that the algorithm for maximizing the function $\Phi'$ with $\lambda = \frac{1}{2}$, which is equivalent to mRMR, does not provide the optimal result.

In contrast, decreasing the value of the parameter $\lambda$ to $\frac{1}{5}$ results in increasing the value of $\Phi'(\{X_1, X_2\})$:

$$\Phi'(\{X_1, X_2\}) = \frac{1}{2}(I(X_1; Y) + I(X_2; Y)) - \frac{1}{10}(I(X_1; X_2) + I(X_2; X_1)) \cong 0.080, \tag{20}$$

and because the value of $\Phi'(\{X_1, X_3\})$ remains unchanged, in this case, the algorithm for maximizing the function $\Phi'$ selects the optimal variable $X_2$ in the second step.

The critical value of the parameter $\lambda$ for this dataset is approximately 0.297, which is obtained by comparing the values of $\Phi'(\{X_1, X_2\})$ and $\Phi'(\{X_1, X_3\})$, i.e., from the equation:

$$\begin{aligned} &I(X_1; Y) + I(X_2; Y) - \lambda(I(X_1; X_2) + I(X_2; X_1)) \\ &\quad = I(X_1; Y) + I(X_3; Y) - \lambda(I(X_1; X_3) + I(X_3; X_1)). \end{aligned} \tag{21}$$

When the value of $\lambda \geqslant 0.297$, the algorithm for maximization of the function $\Phi'$ selects the set $\{X_1, X_3\}$, whereas for the values of $\lambda < 0.297$, the algorithm selects the optimal set $\{X_1, X_2\}$.

We define the objective function $\Phi'$ in (18). Maximization of $\Phi'$ for the parameter $\lambda = \frac{1}{2}$ is equivalent to the mRMR algorithm, for the first-order incremental search. We show that the value of $\lambda$ can influence the quality of the FS.

It is worth noting that for the counterexample shown in Subsection 3.1, there is no positive parameter $\lambda$ of the function $\Phi'$ for which the maximization of the function $\Phi'$ leads to the same solution as the Max-Dependency algorithm. This can be proven by solving (21).

## 4 Comparison of different dependency-based methods: numerical experiments

The mRMR method can be interpreted as a framework for effective FS that provides the potential for novel sophisticated and effective schemes [14]. One possible approach is to select the convenient measure

of dependency that is a crucial aspect in the mRMR methodology [18]. In the case of continuous features, the correlation coefficients as Pearson and Spearman, instead of the mutual information can be used. For discrete variables, one can employ also chi-squared test statistics. Recently, a new correlation measure—distance correlation [28] has been proposed and is gaining popularity. Owing to its characteristics, it can be a suitable measure for use in dependency-based FS methods.

### 4.1 Distance correlation

To define the distance correlation, the distance covariance needs to be defined. Let $(z_i, y_i)$ for $i = 1, 2, \ldots, n$ be statistical samples of two random variables $(Z, Y)$. Variables $Z$ and $Y$ can be multidimensional and have different sizes. Additionally, we define $(a_{i,j})$ and $(b_{i,j})$ as distance matrices of size $n \times n$ with matrix elements:

$$\begin{aligned} a_{i,j} &= L2(z_i, z_j), \quad i, j = 1, 2, \ldots, n, \\ b_{i,j} &= L2(y_i, y_j), \quad i, j = 1, 2, \ldots, n, \end{aligned} \tag{22}$$

where $L2$ stands for the Euclidean distance. Let $A_{i,j}$ and $B_{i,j}$ for $i, j = 1, 2, \ldots, n$ be elements of double centered distance matrices:

$$\begin{aligned} A_{i,j} &= a_{i,j} - \bar{a}_{i.} - \bar{a}_{.j} + \bar{a}_{..}, \quad \text{for } i, j = 1, 2, \ldots, n, \\ B_{i,j} &= b_{i,j} - \bar{b}_{i.} - \bar{b}_{.j} + \bar{b}_{..}, \quad \text{for } i, j = 1, 2, \ldots, n, \end{aligned} \tag{23}$$

where $\bar{a}_{i.}$ is the mean value of the $i$th row, $\bar{a}_{.j}$ is the mean value of the $j$th column and $\bar{a}_{..}$ is the mean value of all matrix elements.

Then, the squared sample distance covariance is defined as the arithmetic average of the products $A_{i,j}B_{i,j}$:

$$\mathrm{dCov}_n^2(Z, Y) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{i,j} B_{i,j}. \tag{24}$$

The distance variance is a special case of distance covariance when two random variables are identical. The squared sample distance variance is defined as

$$\mathrm{dVar}_n^2(Z) = \mathrm{dCov}_n^2(Z, Z). \tag{25}$$

The sample distance correlation of two random variables is determined by the ratio:

$$\mathrm{dCor}_n(Z, Y) = \frac{\mathrm{dCov}_n(Z, Y)}{\sqrt{\mathrm{dVar}_n(Z)\mathrm{dVar}_n(Y)}}. \tag{26}$$

The distance correlation measures dependency between two arbitrary random variables, not necessarily of equal dimension. It is zero if and only if the random variables are independent. Because for the classical correlation measures such as the Pearson and Spearman coefficients, this claim is valid only in one direction, they can easily be zero for dependent variables. Whereas the Pearson (Spearman) correlation coefficient is sensitive mainly to the linear (monotonic) relationship between two random variables, the distance correlation can capture both linear and nonlinear association between them. Therefore, the distance correlation is one of the measures we used in numerical experiments to compare the performance of mRMR and some other FS methods based on dependency measures.

### 4.2 Different dependency-based methods: numerical experiments

We evaluate the performance of the mRMR FS method with various dependency measures. The versions of mRMR are compared with two other FS methods based on the dependency. The first of them is the first-order incremental Max-Dependency (MD) algorithm that looks for a feature subset on which the target variable is most dependent. The second method ranks the features individually according to their dependency on the target variable (Rank), with a higher value feature being more important. In addition to the mutual information (MI) as an original dependency measure for mRMR, other measures

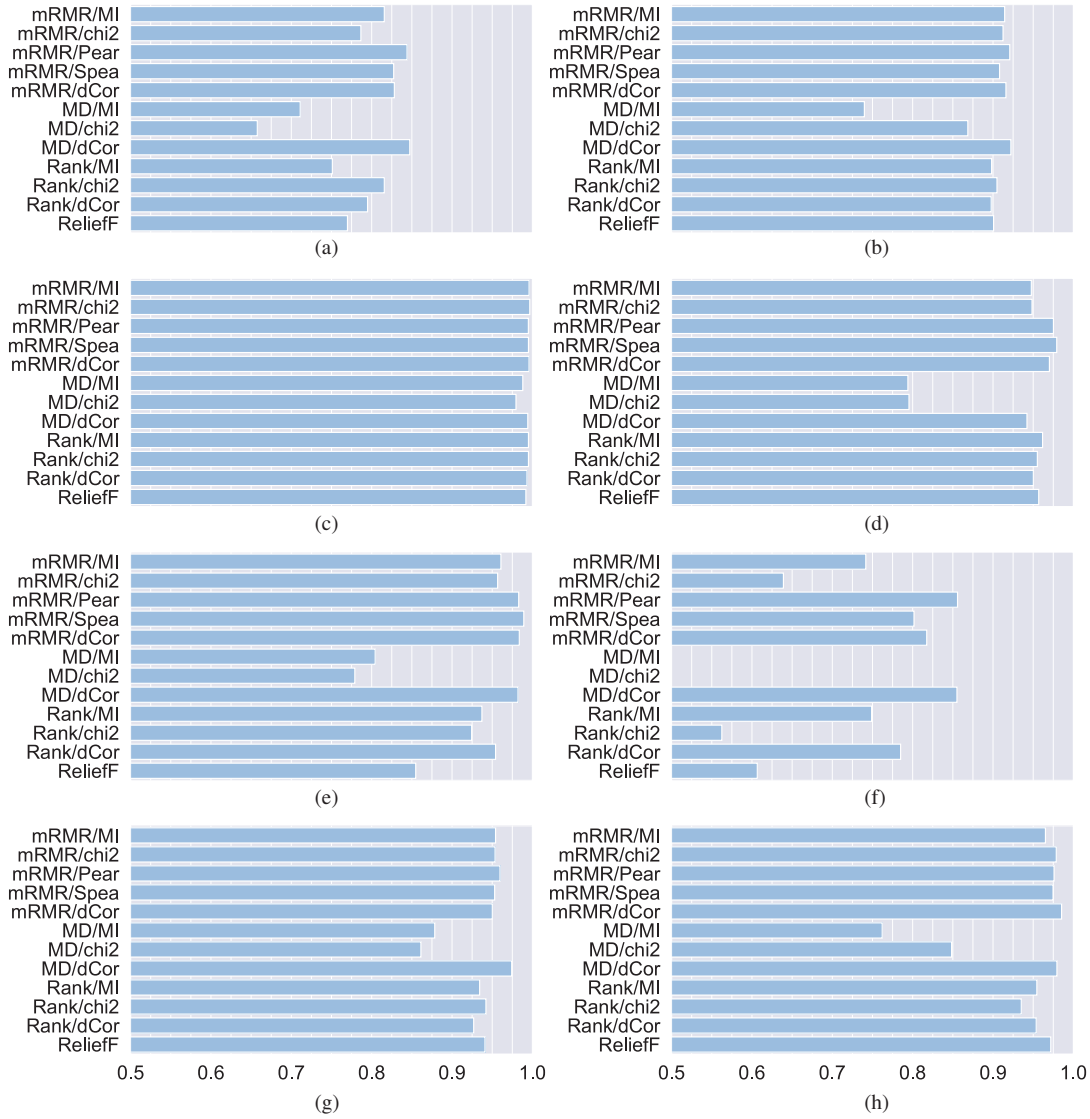


**Figure 1** (Color online) The average $F_1$ score of four classifiers (NB, SVC, RF, kNN) on eight high-dimensional datasets. Comparison of multiple FS approaches. (a) Alon; (b) Tian; (c) Gordon; (d) Golub; (e) Burczynski; (f) Pomeroy; (g) Singh; (h) Chowdary.

such as chi-squared test statistics (chi2), Pearson correlation coefficient (Pear), Spearman correlation coefficient (Spea), and distance correlation (dCor) are used. We also include the state-of-the-art FS method reliefF [29] in the comparison.

For evaluation, we use eight high-dimensional real-world datasets. They are publicly available DNA microarray datasets that constitute the binary classification tasks or are converted to the binary task. Their basic characteristics are provided in Table 3.

To evaluate the influence of the FS methods on the prediction performance, we use four different classifiers, Gaussian naive Bayes classifier, support vector classifier (the penalty parameter $C = 1$), random forest classifier (1000 base estimators, the entropy function to measure the quality of a split), and $k$ nearest neighbors classifier ($k = 5$). Stratified 10-fold cross-validation is used to validate the results.

Because several of the utilized datasets are imbalanced, the accuracy is not an appropriate measure of the prediction performance. Instead of accuracy, we use $F_1$ score to measure the performance of the classifiers.

To provide the compact overview of the classification results, we show the average classification results of four classifiers in Figure 1 and the best results of four classifiers in Figure 2.

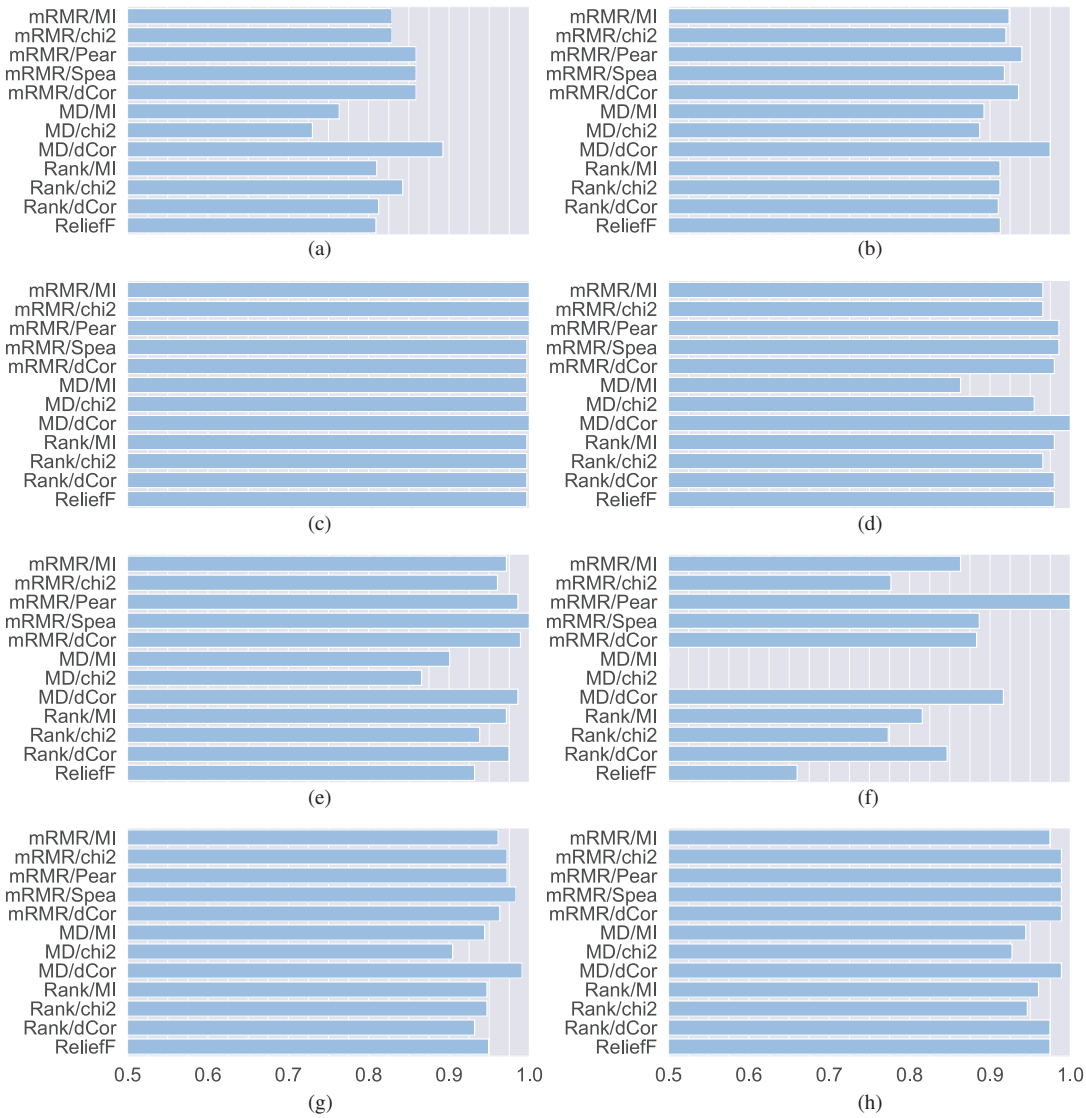


**Figure 2** (Color online) The highest $F_1$ score of four classifiers (NB, SVC, RF, kNN) on eight high-dimensional datasets. Comparison of multiple FS approaches. (a) Alon; (b) Tian; (c) Gordon; (d) Golub; (e) Burczynski; (f) Pomeroy; (g) Singh; (h) Chowdary.

When comparing the results, it can be observed that the methods Max-Dependency using mutual information (MD/MI) and chi-squared test statistics (MD/chi2) give generally worse results than the other evaluated methods. The problem of MD/MI and MD/chi2 is in estimating the probability of possible states from the complete contingency table. When using only five features, each with three values, the complete contingency table has $3^5 \times 2 = 486$ items, what is more than the number of samples in the datasets used. The examined datasets have a too-small number of samples to reliably estimate the probability of possible states and to compute the MI. The results confirm exceptions described in [14] that the MD/MI method is appropriate in selecting only the small number of relevant features and needs a larger number of samples. The same is true for the MD/chi2 method.

The results of the Max-Dependency FS method using distance correlation (MD/dCor) are promising. When comparing the highest $F_1$ score of four classifiers, MD/dCor accomplishes the highest precision of the prediction on all datasets except the Burczynski and Pomeroy datasets and this method significantly outperforms the others. When comparing the average $F_1$ score, the results of MD/dCor and some versions of mRMR are more balanced. The best method of MD/dCor is followed by mRMR using the Spearman correlation coefficient (mRMR/Spea).

**Table 5** Summary of results of prediction performance for different FS methods

| | WTL | Mean rank | Standardized mean | Standardized median |
|---|---|---|---|---|
| mRMR/MI | 183/55/114 | 5.4 | 0.44 | 0.41 |
| mRMR/chi2 | 174/44/134 | 5.9 | 0.33 | 0.41 |
| mRMR/Pear | 249/51/52 | 3.4 | 0.7 | 0.73 |
| mRMR/Spea | 222/46/84 | 4.3 | 0.59 | 0.59 |
| mRMR/dCor | 234/53/65 | 3.9 | 0.63 | 0.63 |
| MD/MI | 33/9/310 | 10.8 | −1.54 | −1.53 |
| MD/chi2 | 19/12/321 | 11.2 | −1.86 | −1.97 |
| MD/dCor | 249/28/75 | 3.8 | 0.64 | 0.81 |
| Rank/MI | 146/28/178 | 7.0 | 0.11 | 0.26 |
| Rank/chi2 | 130/38/184 | 7.4 | 0.04 | 0.17 |
| Rank/dCor | 134/38/180 | 7.2 | 0.12 | 0.25 |
| ReliefF | 123/30/199 | 7.7 | −0.21 | 0.01 |

To support comparative analysis of utilized FS methods, we provide Table 5 showing Winn/Tie/Loss (WTL) statistics, mean rank, standardized mean, and standardized median. We compare all FS methods on the eight datasets from Table 3 using four classifiers. These results show that the best performing methods are mRMR/Pear, MD/dCor, mRMR/dCor, and mRMR/Spea. In contrast with the results achieved by MD/dCor, the results of MD/MI and MD/chi2 are yielding poor prediction performance.

When analyzing the mRMR results, although the Pearson and Spearman correlation coefficients reveal only linear or monotonic dependencies, they may be useful when mRMR is applied to continuous data. The transformation of continuous to discrete data for using MI or chi-squared statistics leads to loss of information and thus worsens the results of mRMR/MI and mRMR/chi2.

The experiments confirm that the distance correlation is a suitable measure for such tasks. If we compare all the three approaches to use distance correlation—simple ranking (Rank), mRMR, and Max-Dependency, ranking appears to be the worst, whereas Max-Dependency achieves the best results, under expectations. The weakness of the methods using distance correlation is their worse usability in the case of a larger number of observations because of the need to calculate a distance matrix with the quadratic complexity.

### 4.3 Generalization of mRMR algorithm: numerical experiments

In Subsection 3.2, we define the objective function $\Phi'$ in (18) maximization of which for the parameter $\lambda = \frac{1}{2}$ is equivalent to the mRMR algorithm in MID version, for the first-order incremental search.

To examine the influence of the $\lambda$ parameter on FS quality, we conduct experiments with various values of $\lambda$ for eight real-world datasets stated in Table 3. We use nine values of $\lambda = 0.00$, 0.25, 0.50 up to 2.00. The value of $\lambda$ determines a weight of the average redundancy in the objective function $\Phi'$ in (18). If the weight is $\lambda = 0.00$, redundancy is not considered and the optimization task is equivalent to the Rank/MI method. For $\lambda = 0.50$, the average redundancy has half the weight of the average relevance and the optimization task is equivalent to mRMR. As the maximum $\lambda$, we use the value of 2.00, where the weight of the average redundancy is twice as the weight of the average relevance.

The experiments show that the value of $\lambda$ influences the quality of the FS. For all datasets, we obtain distinct feature subsets for all $\lambda$ values. The optimal value of this parameter depends on a particular dataset, e.g., in the case of the Pomeroy dataset, maximum $F_1$ score is achieved with $\lambda = 0.25$, whereas for the Singh dataset, it is for $\lambda = 1.75$.

The summary of the experimental results is in Table 6, where we present the average ranking obtained on examined datasets and computed from maximum and average $F_1$ score for all four used classifiers. As the results show, too-small values of $\lambda$ (0.00, 0.25) or too large values (2.00) yield sub-optimal results. For the test datasets in general, the optimal values are in the range 0.75 to 1.50, which are higher weights of redundancy as in the standard mRMR.

**Table 6** Generalization of mRMR. Mean rank of maximum and average $F_1$ score versus parameter $\lambda$

| Parameter $\lambda$ | 0.00 | 0.25 | 0.50 | 0.75 | 1.00 | 1.25 | 1.50 | 1.75 | 2.00 |
|---|---|---|---|---|---|---|---|---|---|
| $F_1$ score MAX | 7.25 | 5.81 | 5.44 | **4.19** | **4.38** | **4.69** | **3.63** | **4.06** | 5.56 |
| $F_1$ score AVG | 6.88 | 5.38 | **4.50** | **3.81** | **4.06** | **3.94** | **4.69** | 5.13 | 6.63 |

## 5 Conclusion

FS is one of the key concepts in machine learning which has a significant impact on the performance and interpretability of a predictive model. In this paper, we focused on the minimum redundancy maximum relevance (mRMR) feature selection method, which is the very popular FS method based on the information theory. We analyzed the relationship between mRMR and Max-Dependency criterion and constructed the discrete counterexample to the theorem about the equivalence of mRMR and Max-Dependency for the first-order incremental search presented by the authors of the mRMR method. We also showed the examples of real-world datasets on which the mRMR method and the incremental Max-Dependency algorithm give different results. In addition, we defined the objective function whose maximization is equivalent to mRMR and provided the generalization of the mRMR method. In the numerical experiments, we compared three dependency-based approaches, mRMR, Max-Dependency, and simple ranking, using various dependency measures in terms of their influence on the prediction accuracy of classifiers. The results on eight high-dimensional real-world datasets showed that the performance of mRMR originally used with MI can be improved by employing the other suitable dependency measures. When comparing the maximal performance of classifiers used, the Max-Dependency incremental algorithm with distance correlation as a measure outperformed the other presented FS methods and looked as a promising FS approach for the case of small sample sizes problems frequently occurring in biomedical research.

**Acknowledgements** This work was supported by Slovak Research and Development Agency (Grant No. APVV-16-0211).

**References**

1 Li Y, Li T, Liu H. Recent advances in feature selection and its applications. Knowl Inf Syst, 2017, 53: 551–577
2 Li J D, Liu H. Challenges of feature selection for big data analytics. IEEE Intell Syst, 2017, 32: 9–15
3 Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A. Recent advances and emerging challenges of feature selection in the context of big data. Knowledge-Based Syst, 2015, 86: 33–45
4 Ang J C, Mirzal A, Haron H, et al. Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. IEEE/ACM Trans Comput Biol Bioinf, 2016, 13: 971–989
5 Li J D, Cheng K W, Wang S H, et al. Feature selection. ACM Comput Surv, 2018, 50: 1–45
6 Battiti R. Using mutual information for selecting features in supervised neural net learning. IEEE Trans Neural Netw, 1994, 5: 537–550
7 Kwak N, Choi C H. Input feature selection for classification problems. IEEE Trans Neural Netw, 2002, 13: 143–159
8 Cai R C, Hao Z F, Yang X W, et al. An efficient gene selection algorithm based on mutual information. Neurocomputing, 2009, 72: 991–999
9 Fleuret F. Fast binary feature selection with conditional mutual information. J Mach Learn Res, 2004, 5: 1531–1555
10 Cheng H R, Qin Z G, Feng C S, et al. Conditional mutual information-based feature selection analyzing for synergy and redundancy. ETRI J, 2011, 33: 210–218
11 Yang H H, Moody J. Data visualization and feature selection: new algorithms for nongaussian data. In: Proceedings of the 12th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 1999. 687–693
12 Vergara J R, Estévez P A. A review of feature selection methods based on mutual information. Neural Comput Appl, 2014, 24: 175–186
13 Brown G, Pocock A, M-Zhao J, et al. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. Mach J Learn Res, 2012, 13, 27–66
14 Peng H C, Long F H, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Machine Intell, 2005, 27: 1226–1238
15 Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. J Bioinform Comput Biol, 2005, 03: 185–205
16 Corredor G, Wang X, Zhou Y, et al. Spatial architecture and arrangement of tumor-infiltrating lymphocytes for predicting likelihood of recurrence in early-stage non-small cell lung cancer. Clin Cancer Res, 2019, 25: 1526–1534
17 Toyoda A, Ogawa T, Haseyama M. Favorite video estimation based on multiview feature integration via KMvLFDA. IEEE Access, 2018, 6: 63833–63842

18 Berrendero J R, Cuevas A, Torrecilla J L. The mRMR variable selection method: a comparative study for functional data. J Stat Comput Simul, 2016, 86: 891–907
19 Guyon I, Elisseeff A. An introduction to variable and feature selection. Mach J Learn Res, 2003, 3: 1157–1182
20 Golub T R, Slonim D K, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 1999, 286: 531–537
21 Gordon G J G, Jensen R V R, Hsiao L-L L, et al. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Res, 2002, 62: 4963–4967
22 Alon U, Barkai N, Notterman D A, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA, 1999, 96: 6745–6750
23 Tian E, Zhan F, Walker R, et al. The role of the Wnt-signaling antagonist DKK1 in the development of osteolytic lesions in multiple myeloma. New Engl J Med, 2003, 349: 2483–2494
24 Burczynski M E, Peterson R L, Twine N C, et al. Molecular classification of crohn's disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells. J Mol Diagn, 2006, 8: 51–61
25 Pomeroy S L, Tamayo P, Gaasenbeek M, et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature, 2002, 415: 436–442
26 Singh Y N, Singh S K, Ray A K. Bioelectrical signals as emerging biometrics: issues and challenges. ISRN Signal Process, 2012, 2012: 136–151
27 Chowdary D, Lathrop J, Skelton J, et al. Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative. J Mol Diagn, 2006, 8: 31–39
28 Székely G J, Rizzo M L, Bakirov N K. Measuring and testing dependence by correlation of distances. Ann Statist, 2007, 35: 2769–2794
29 Robnik-Šikonja M, Kononenko I. Theoretical and empirical analysis of ReliefF and RReliefF. Mach Learn, 2003, 53: 23–69