Maximum-Relevance and Maximum-Complementarity Feature Selection with Random Forest

Mudan Zhou[1], Pei Liu[2], and Fan Yang[3](✉)

[1] School of Information Science and Technology, Xiamen University
Tan Kah Kee College, Xiamen, China
mudanzhou@xujc.com

[2] School of Aerospace Engineering, Xiamen University, Xiamen, China
peiliu@stu.xmu.edu.cn, yang@xmu.edu.cn

[3] Shenzhen Research Institute, Xiamen University, Shenzhen, China

Abstract. In feature selection for high-dimensional data, in order to select the minimum number of features that can well explain the target, it requires finding the relevant features to the predictive target, as well as removing the redundant information by discovering the feature interactions. Existing approaches usually measure the feature interactions using relevance between features without considering their joint dependencies on the target. In this paper, we propose a new feature selection criterion, Maximum-Relevance and Maximum-Complementarity (MRMC). Besides the relevance with the target, MRMC takes into consideration the complementary information of a candidate feature to a selected feature or feature subset when predicting the target. We then present an efficient approach to calculate the information complementarity between features with random forests. Finally we implement MRMC Feature Selection using sequential forward search (SFS). Experimental results on 18 data sets show that SFS-MRMC achieved the best overall performances compared with other information-theoretical feature selection methods and RF-RFE.

Keywords: Feature selection · Information entropy · Random forest

1 Introduction

Feature selection is of crucial importance for high-dimensional data classification. It can help reduce the size of the feature space, thereby reduce the computational cost and time complexity of the learning model. Moreover, it aims to identify the smallest subset of features that can explain the target, thus it can enhance the model interpretability.

Information entropy based feature selection methods are widely used. They are easy to interpret and implement, due to their information theoretical framework. These approaches can always be classified as filter methods. They are

Z. Ren et al. (Eds.): *Proceedings of 2021 5th Chinese Conference on Swarm Intelligence and Cooperative Control*, LNEE 934, pp. 503–514, 2023.
https://doi.org/10.1007/978-981-19-3998-3_48

independent of the subsequent learning algorithm, which facility reducing the computational cost. Information entropy based methods generally use information criteria to measure the relevance between features and target classes, and the interaction within features. Mutual information is widely used to measure the relevance between two variables in information entropy based feature selection. For example, the Relevance criterion (REL) [1] considers the relevance between features and the targets, while Maximum relevance and minimum redundancy (mRMR) [2] and CIFE [3] utilizes mutual information to measure the relevance between features and target class and the redundancy between features. However, most existing approaches usually measure the feature interactions only using correlation between features without considering their joint dependencies on the target. That is to say, they ignore the information of classification complementarity among the features. In contrast, the wrapper method and embedded method, such as variable importance measures in random forests [4,5], consider the feature interactions conditioned on the target classes. They always obtain more satisfactory classification results along with the learning algorithm.

In this paper, we propose a new feature selection method, Maximum-Relevance and Maximum-Complementarity (MRMC). The main contributions are as follows.

1. We give the formal definition of information complementarity in the information-theoretical framework. Besides the relevance with the target, MRMC takes into consideration the complementary information of a candidate feature providing to a selected feature or a selected feature subset to predict the target.
2. To efficiently measure the information complementarity conditioned on the target classes, we present an efficient approach to approximate the complementarity between pairs of features with random forests. Thus it can be viewed as a hybrid feature selection approach with takes the advantages of both entropy based method and random forest.
3. Experiments results on 18 datasets in comparison with classical and state-of-the-art approaches, i.e. CMIM [15], mRMR [2], DISR [16], JMI [17] and RF-RFE [11], demonstrate the effectiveness and efficiency of our method.

The paper is structured as follows: in Sect. 2 we states the related work. In Sect. 3, we describe the proposed feature selection algorithm. Extensive experimental results are provided in Sect. 4. In Sect. 5, we summarize our work.

2 The Information-Theoretical Framework for Feature Selection

In this section we review the related works on information entropy based feature selection. To understand them, we first introduce some basic concepts of relevant features and redundant features from the perspective of information theory.

Definition 1. *Given discrete random variable X and its probability distribution $p(x) = P(X = x)$ with domain X, the entropy of random variable X is defined as:*

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \tag{1}$$

$H(X)$ indicates the amount of information needed to eliminate its uncertainty, that is, the amount of information that X may contain.

Definition 2. *Given discrete random variables X and Y with domain X and Y and their joint probability distribution $p(x, y) = P(X = x, Y = y)$, then the conditional entropy of random variable Y given X is defined as:*

$$H(Y|X) = -\sum_{y \in Y} \sum_{x \in X} p(x, y) \log_2 p(y|x) \tag{2}$$

In holds that $0 \leq H(Y|X) \leq H(Y)$. $H(Y|X) = 0$ with the condition that $p(y|x) = 1$ or $p(y|x) = 0$ for any pair x, y, in simple words, the value of Y is determined given the value of X. $H(Y|X) = H(Y)$ with X and Y are independent.

Definition 3. *Given discrete random variables X and Y, then their mutual information is defined as:*

$$I(X; Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \tag{3}$$

$I(X, Y)$ can be interpreted as the amount of reduced uncertainty of Y due to X. Therefore, $I(X, Y)$ represents the relevance between X and Y.

Definition 4. *Given random variables X, Y and Z, then the conditional mutual information of X and Y given Z is defined as:*

$$\begin{aligned} I(X; Y|Z) &= H(Y|Z) - H(Y|Z, X) \\ &= H(X|Z) - H(X|Z, Y) = I(Y; X|Z) \end{aligned} \tag{4}$$

The conditional mutual information quantifies the reduction of uncertainty of Y(X) owing to the variable $X(Y)$ given Z is known.

Definition 5 *(Chain rule for mutual information)*

Given a set of random variables $X_S = \{X_1, X_2, \cdots, X_n\}$ and random variable Y, then the mutual information of X_S and Y is defined as:

$$\begin{aligned} I(X_S; Y) &= I(X_1, X_2, \cdots, X_n; Y) \\ &= \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, X_{i-2}, \cdots, X_1) \end{aligned} \tag{5}$$

The chain rule for mutual information indicates the amount of information that the random variables set X_S can provide for Y equals to the sum of pairwise mutual information of Y and each variable under certain conditions. This is an important trick which has been used in many feature selection methods when considering the influence a feature subset on the target class.

2.1 The Relevant Information of Features

In feature selection, the mutual information $I(X_S; Y)$ can be used to measure the dependency between input feature subset X_S and output target Y. Suppose X_k be a candidate feature and $X_S \subset \{X_1, \cdots, X_{k-1}, X_{k+1}, \cdots, X_n\}$ be a feature subset selected, the relevance of X_k is calculates as:

$$I(X_k; Y|X_S) = I(\{X_S, X_k\}; Y) - I(X_S; Y) \tag{6}$$

It measures how much the candidate feature X_k is relevant to Y when X_S is given. From it, we can observe that the relevance in feature selection is conditional, as pointed out by [18,19]. For a candidate feature, it may be strongly relevant, weakly relevant or irrelevant when conditioned by different context X_S.

The Relevance criterion (REL) use $I(X_k; Y|X_S)$ combined with forward feature selection directly, known as maximal dependency [1]. That is in each step the feature X_k is selected which meets:

$$J_{REL}(X_k) = \max_{X_k \in X_{-S}} I(X_k; Y|X_S) \tag{7}$$

where $X_{-S} = X/X_S$. There is a major drawback that we need to estimate multivariate probability density to measure the feature subset-conditional mutual information. It is almost impossible to estimate with limited data if features are all correlated to each other. To solve those problem, the bi-variate or tri-variate probability density is employed in existing methods [2,20]. For example, according to Conditional Mutual Information Maximization criterion (CMIM) [15], in each step, the candidate feature can be selected as follows:

$$J_{CMIM}(X_k) = \max_{X_k \in X_{-S}} \min_{X_j \in X_S} I(X_k; Y|X_j) \tag{8}$$

It can seen that CMIM uses $\min_{X_j \in X_S} I(X_k; Y|X_j)$ to replace $I(X_k; Y|X_S)$, because the latter is hard to compute when the selected feature subset increases.

2.2 The Redundant Information Within Features

Neglecting the feature interactions may cause redundancy in the selected feature subset. Some methods focus on removing feature redundancy when taking into account feature relevance with predictive targets [19,21,22]. For example, it can be computed through mutual information between the candidate feature and the selected feature subset as follows,

$$R(X_S; X_k) = \alpha \sum_{X_j \in X_S} I(X_j; X_k) \tag{9}$$

where X_S is a feature subset selected, X_k is a candidate feature. This measurement of redundancy has been applied in mRMR [2]. Its objective is as follows:

$$J_{MRMR}(X_k) = I(X_k; Y) - \frac{1}{|X_S|} \sum_{X_j \in X_S} I(X_k; X_j) \tag{10}$$

It considers the mutual information between features as redundant information. However the target class Y is neglected in measuring the redundancy. Figure 1 gives a intuitive example. Two shadow parts in Fig. 1 represent $I(X_1; X_2)$, but only red shadow part kicks in classifying the target class Y. Therefore, some methods, e.g. CIFE [23], MIFS-U [24], mIMR [25] and IGFS [26] all use joint mutual information $I(X_i; X_k; Y)$ to measure the feature redundancy, which can defined as follows:

$$I(X_i; X_k; Y) = I(X_i; Y) + I(X_k; Y) - I(X_i, X_k; Y) \tag{11}$$

Similarly, given a selected feature subset X_S, the joint mutual information $I(X_S; X_k; Y)$ measures the feature redundancy of the candidate feature X_k when given X_S and Y. In this sense, the joint mutual information can be regarded as the shared discriminative information of $\{X_S, X_k\}$ about Y. However, similarly, it is hard to compute $I(X_S; X_k; Y)$, so the sum of pairwise redundancies between features is always calculated to approximate it. For example, the criterion of $CIFE$ is defined as:

$$J_{CIFE}(X_k) = I(X_k; Y) - \sum_{X_j \in X_S} I(X_k; X_j; Y) \tag{12}$$

In [19,27], the markov blanket was also used to evaluate the redundancy between features.

3 Maximum-Relevance and Maximum-Complementarity Feature Selection

In this section, we first give the definition of feature complementarity. Two features with more complementarity indicates less redundancy. Then we give the objective of MRMC and propose to use random forests to approximate the complementarity score between any pair of features. Finally, for simplicity, a sequential forward search strategy is employed to maximize the objective function.

3.1 The Complementary Information of Features

Besides relevancy and redundancy, there are some literature have proposed the definition of complementarity from different aspect. Paper [16] proposes that complementarity is the beneficial effect if feature interaction. From this part, the complementary information is the negative interaction information. The definition is as follows,

Definition 6. *Suppose $X_1, X_2, \cdots, X_n$ are random variables, they are complementary if*

$$(-1)^n I(X_1; X_2; \cdots; n) > 0$$

That is, for two variables, the classification information about target provided by two features together is greater than the sum of the classification information about provided by two features individual. If $I(X_i, X_j; Y) > I(X_i; Y) + I(X_j; Y)$, the features X_i and X_j are said to be complementary.

However, it is hard to compute the complementarity of a large feature set. In [28], the authors proposed JMI, which defines the complementary information as the shared information between two features X_i and X_j given the target, i.e. $I(X_i, X_j; Y)$. Thus they defined complementary information of X_k provided to the selected subset X_S as follows,

$$J_{JMI}(X_k) = \sum_{X_j \in X_S} I(X_k, X_j; Y)$$

In [29], the relevance and complementary score are estimated by using neural network, which is highly time-consuming.

Unlike the negative interaction information and shared information, we formally give the definition of feature complementarity from the perspective of information entropy which is the sum of the unique information of two features about target.

Definition 7. *Suppose X_1 and X_2 are two random variables used for predicting the target variable Y, their complementary classification information between two variables (CCI) can be defined as follows,*

$$CCI(X_1, X_2; Y) = \frac{I(X_1; Y|X_2) + I(X_2; Y|X_1)}{2} \tag{13}$$

Definition 8. *Suppose X_k be a random variable and $X_S = \{X_1, X_2, \cdots, X_{k-1}\}$ be a random variable subset, their complementary classification information (CCI) for predicting the target variable Y can be defined as follows,*

$$CCI(X_k, X_S; Y) = \frac{I(X_k; Y|X_S) + I(X_S; Y|X_k)}{|X_S| + 1} \tag{14}$$

CCI quantifies the new classification information provided by a feature when another feature or subset is given. Suppose X_k is a candidate feature and X_S is selected feature subset, $I(X_k; Y|X_S)$ measures the amount of newly provided classification information by the candidate feature X_k while $I(X_S; Y|X_k)$ indicates the amount of classification information preserved by the selected feature set when the candidate feature is added. Hence CCI measures the classification information of subset $\{X_k, X_S\}$.

It is hard to compute the complementary information between feature and a feature subset. We can also use sum of pairwise CCI between features to approximate it. Assume that dataset $X = \{X_1, X_2, \cdots, X_n\}$ characterized as an n-dimensional vector and X is labeled with L classes $Y = \{y_j\}, j = 1, 2, \cdots, L$. According to the definition of complementarity between features, we can obtain the complementary matrix between any pair of features as follows:

$$C = \{c_{i,j}\}_{1 \leq i,j \leq n} = \begin{bmatrix} 0 & c_{12} & \cdots & c_{1n} \\ c_{21} & 0 & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & 0 \end{bmatrix} \tag{15}$$

where $c_{i,j} = CCI(X_i, X_j; Y)$, C carries all complementary information between any pair of features. According to the definitions of CCI, C is a real symmetrical matrix.

3.2 Maximum-Relevance and Maximum-Complementarity Based Feature Selection

For high-dimensional data, it is infeasible to compute the joint mutual information of many variables. Hence we use the pairwise complementarity of features to help find the optimal feature subset. Suppose X_k is a candidate feature and $X_S = \{X_1, X_2, \cdots, X_{k-1}\}$ is selected subset, the criterion of $MRMC$ can be defined as follows,

$$J_{MRMC}(X_k) = I(X_k; Y) + CCI(X_k, X_S; Y) \tag{16}$$

For simplicity, the criterion can be rewrite as follows,

$$J_{MRMC}(X_k) = I(X_k; Y) + \frac{1}{|X_S|} \sum_{X_j \in X_S} CCI(X_k, X_j; Y) \tag{17}$$

where $|X_S|$ is the size of the current selected feature set.

3.3 Mining the Feature Complementarity with Random Forest

To efficiently estimate the feature complementarity, we propose to calculate the complementary scores between any pair of features using random forest. Inspired by the sample proximity matrix of random forest, we can also get the pairwise complementarity matrix from random forests. The main assumption is that when two features are used in the same tree, then they can be viewed as complementary to each other, since a tree model naturally performs feature selection when it is being built. The frequency of two features being used in the forest can naturally used as an estimate of the feature complementarity.

Given a trained random forest $\{h_1, h_2, \cdots, h_{n_{tree}}\}$, the complementarity score of any pair of features is defined as:

$$\tilde{c}_{i,j} = av_k Idt(X_i, X_j \in h_k), k = 1, 2, \cdots, n_{tree} \tag{18}$$

where n_{tree} is the number of trees in the forest, $Idt(\cdot)$ is the indicator function, and $av(\cdot)$ is the mean operator. Its value is normalized to range $[0, 1]$. A value 0 of $\tilde{c}_{i,j}$represents that there is no additional knowledge to predict Y when using $\{X_i, X_j\}$ compared with only using X_i or X_j. And value 1 reveals that X_i and

X_j have the largest complementarity. The performance of combining X_i and X_j is much better than only employing X_i or X_j.

An efficient greedy search method, Sequential Forward Search (SFS) was employed to obtain a feature subset for MRMC. SFS-MRMC consists the following steps: 1) It starts from an empty feature subset and the relevance scores of all features is calculated. The most relevant feature is selected at first; 2) At each selection step, it expands feature subset with the feature with largest MRMC score.; 3) Repeat step 2) until the stop conditions reached.

4 Experiment and Discussion

4.1 Datasets and Algorithms for Comparison

To evaluate the performance of SFS-MRMC, 18 datasets from different domains are selected from http://featureselection.asu.edu/datasets.php, https://archive.ics.uci.edu/ml/index.php and http://www.gems-system.org/. The description of the datasets is summarized in Table 1. The number of features ranges from 44 to 10000 with categories varying from 2 to 15.

Table 1. Description of datasets

Data set	Features	Instances	Classes	Data set	Features	Instances	Classes
Hearts	44	267	2	BASEHOCK	4862	1993	2
p_gene	57	106	2	DLBCL	5470	77	2
Sonar	60	165	15	Brain_Tumor1	5921	90	5
CHART	60	600	2	Prostate_GE	5966	102	2
Colon	2000	62	2	Leukemia	7070	72	2
SRBCT	2309	83	4	ALLAML	7129	72	2
warpPIE10P	2420	210	10	Central	7129	60	2
Lung	3312	203	5	Carcinom	9182	174	11
Lymphoma	4026	96	9	Arcene	10000	200	2

SFS-MRMC is compared with one Relevance and Redundancy based methods mRMR, three Complementarity based methods CMIM, DISR and JMI, and an embedded method RF-RFE. Note that RF-RFE always achieves state-of-the-art performances on high-dimensional data.

For fair comparison, in RF-RFE and SFS-MRMC, the number of trees in the forest is set to 1000, the number of splitting features per node is set to the default value $m_{try} = \sqrt{p}$, where p is the total number of features of the dataset.

RF-RFE starts from the total set of features, prunes the least important feature from the current feature subset and then retrains a random forest to update the feature ranking at every iteration, so it is very time-consuming for high-dimensional data. To accelerate the experimental process, when the size of current feature subset is larger than 200, the least important 20 percent of features will be removed at each iteration. When the size of current feature subset reaches 200, the least important one will be eliminated at each iteration.

4.2 Evaluation Metrics

To evaluate the performances of different algorithms, the ten-fold cross-validation (10-CV) is conducted for five times on each dataset. In each fold, every algorithm is performed to obtain a feature ranking on the training set respectively. Then the performances of the feature subset selected according to the feature ranking are evaluated on the test set using the following metrics.

Note that our goal is to select a small feature subset for classification, so we only record the results on the top 200 features selected by different algorithms. For datasets with number of features smaller than 200, we record the results on all sizes of selected feature sets, i.e. $1 \sim m$.

Average Test Accuracy. The classical classifier KNN is used to test the average performances of the selected features using different feature selection algorithms in the five runs of 10-CV. For fair comparison, we set the same parameter settings of the classifiers. For KNN, the parameter k is set to 3.

Average Size of Optimal Feature Subset. We aim to get a small feature subset. For different classifiers, we find the optimal feature subset with the highest accuracy on each dataset. Then the proportion of the subset to the total number of features is recorded, and is averaged over the five runs of 10-CV.

All the experiments are conducted on a PC with Intel CPU 8 GB RAM. We use Python 2.7 for coding, as well as Scikit-Feature feature selection repository.

4.3 Results and Discussion

In this section, we present comparison of SFS-MRMC against other feature selection algorithms in terms of the following aspects.

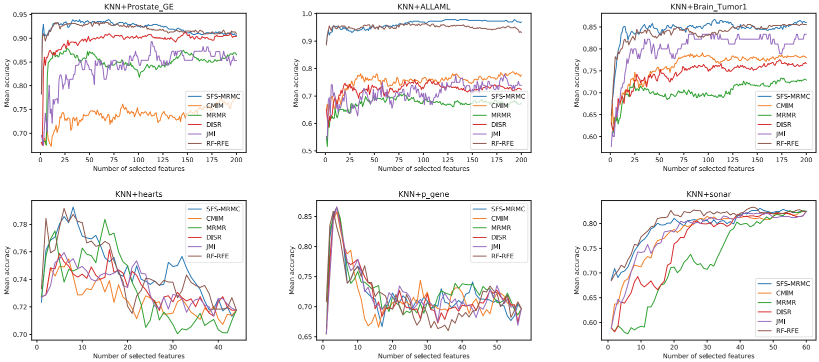

Fig. 1. Accuracy comparison among different feature selection methods

Table 2. The performances of KNN using different feature selection algorithms

Data set	SFS-MRMC	CMIM	MRMR	DISR	JMI	RF-RFE
Hearts	0.7925 ± 0.0141	0.7549 ± 0.0100	0.7836 ± 0.0066	0.7615 ± 0.0172	0.7595 ± 0.0195	0.7915 ± 0.0215
p_gene	0.8658 ± 0.0050	0.8578 ± 0.0151	0.8658 ± 0.0050	0.8658 ± 0.0050	0.8658 ± 0.0050	0.8587 ± 0.0224
Sonar	0.8307 ± 0.0215	0.8153 ± 0.0082	0.8153 ± 0.0125	0.8202 ± 0.0122	0.8221 ± 0.0179	0.8335 ± 0.0168
CHART	0.9880 ± 0.0016	0.9627 ± 0.0023	0.9677 ± 0.0023	0.9670 ± 0.0022	0.9667 ± 0.0045	0.9850 ± 0.0024
Colon	0.8590 ± 0.0188	0.8619 ± 0.0232	0.8486 ± 0.0132	0.8524 ± 0.0197	0.8548 ± 0.023	0.8519 ± 0.0230
SRBCT	0.9950 ± 0.0061	0.8186 ± 0.0104	0.8428 ± 0.0239	0.7961 ± 0.0279	0.7672 ± 0.0378	1.0000 ± 0.0000
warpPIE10P	0.9629 ± 0.0070	0.9333 ± 0.0257	0.9781 ± 0.0023	0.9486 ± 0.0076	0.9286 ± 0.0186	0.9781 ± 0.0038
Lung	0.9597 ± 0.0048	0.9381 ± 0.0095	0.9341 ± 0.0067	0.9419 ± 0.0147	0.9340 ± 0.0066	0.9527 ± 0.0039
Lymphoma	0.8907 ± 0.0138	0.9260 ± 0.0111	0.9218 ± 0.0169	0.9218 ± 0.0103	0.9220 ± 0.0102	0.9216 ± 0.0120
BASEHOCK	0.9287 ± 0.0029	0.9344 ± 0.0017	0.9396 ± 0.0024	0.9448 ± 0.0028	0.9267 ± 0.0026	0.9004 ± 0.0032
DLBCL	0.9486 ± 0.0201	0.8168 ± 0.0257	0.7957 ± 0.0261	0.8107 ± 0.0206	0.8400 ± 0.0370	0.9582 ± 0.0156
Brain_Tumor1	0.8667 ± 0.0199	0.7911 ± 0.0109	0.7333 ± 0.0141	0.7756 ± 0.0285	0.7956 ± 0.0269	0.8578 ± 0.0163
Prostate_GE	0.9393 ± 0.0034	0.7755 ± 0.0340	0.8791 ± 0.0209	0.9116 ± 0.0108	0.8845 ± 0.0187	0.9333 ± 0.0108
Leukemia	0.9807 ± 0.0118	0.9807 ± 0.0069	0.9836 ± 0.0061	0.9826 ± 0.0061	0.9864 ± 0.0009	0.9864 ± 0.0009
ALLAML	0.9782 ± 0.0069	0.7882 ± 0.0126	0.7082 ± 0.0201	0.7604 ± 0.0358	0.7675 ± 0.0328	0.9654 ± 0.0111
Central	0.6333 ± 0.0365	0.6200 ± 0.0694	0.5867 ± 0.0267	0.5867 ± 0.0267	0.6000 ± 0.0483	0.6267 ± 0.0389
Carcinom	0.9108 ± 0.0041	0.7718 ± 0.0296	0.8520 ± 0.0172	0.7983 ± 0.0278	0.7567 ± 0.0136	0.9219 ± 0.0116
Arcene	0.8800 ± 0.0100	0.7440 ± 0.0146	0.6580 ± 0.0150	0.8550 ± 0.0089	0.8390 ± 0.0080	0.8710 ± 0.0107
Average	0.9006 ± 0.0116	0.8384 ± 0.0178	0.8386 ± 0.0132	0.8501 ± 0.0158	0.8454 ± 0.0184	0.8997 ± 0.0125
Average rank	1.8889	3.9444	3.9444	3.6111	3.8889	2.1667

Table 3. Average size of optimal feature subsets selected with different methods (%)

Method	SFS-MRMC	CMIM	MRMR	DISR	JMI	RF-RFE
KNN	12.7961	12.6372	13.8172	13.9300	11.9900	11.1372

Average Test Accuracy. Limited by length, we only show the results of the average test accuracy of KNN on six datasets with different sizes of feature sets selected by different feature selection algorithms in Fig. 1. The number of selected features ranges from 1 to 200. Obviously, different datasets exhibits different variations in terms of test accuracy. For some data sets, such as Prostate_GE, ALLAML, Brain_Tumor1, all feature selection algorithms reach their best performance only with a small number of features. In contrast, for some datasets, such as sonar, the test accuracy exhibits a rising trend with the growth of selected features. And for other datasets, such as hearts and p_gene, more features may deteriorate the performance. We can observe that in most cases SFS-MRMC and RF-RFE are significantly better than other algorithms, while the performance of SFS-MRMC is slightly better than RF-RFE.

Table 2 shows the mean and standard deviations of the best test accuracy of KNN classifier over 5 runs of 10-CV, with different feature selection algorithms respectively. The numbers in red indicate the largest value of each row, i.e. the best results on each datasets, while the numbers in blue indicate the second best accuracy on the datasets. In the bottom, we also display the average performances in terms of accuracy and ranks of each algorithm over the datasets, which show that SFS-MRMC and RF-RFE outperform other algorithms. For all the three classifiers, SFS-MRMC always achieves the highest average ranks.

Average Size of Optimal Feature Subset. Table 3 records the average proportion of selected features over 18 data sets of the six feature selection algorithms with KNN classifiers respectively. Generally, all the six feature selection algorithms achieve remarkable reduction of feature dimension by only selecting a small proportion of the original feature sets.

5 Conclusions

MRMC takes into consideration both the complementarity and relevance within features. We approximate the complementarity between pairs of features using random forests. Thus it can be viewed as a hybrid feature selection approach of both entropy based method and random forest. Experiments results show that SFS-MRMC outperforms four classical entropy based methods and state-of-the-art RF-RFE in terms of classification accuracy, while it is efficient in terms of time cost.

Acknowledgement. This work is supported by the Shenzhen Fundamental Research Program under Grant No. JCYJ20190809161603551.

References

1. Bell, D.A., Wang, H.: A formalism for relevance and its application in feature subset selection. Mach. Learn. **41**(2), 175–195 (2000)
2. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. **27**(8), 1226–1238 (2005)
3. Vergara, J.R., Estévez, P.A.: A review of feature selection methods based on mutual information. Neural Comput. Appl. **24**(1), 175–186 (2013). https://doi.org/10.1007/s00521-013-1368-0
4. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
5. Genuer, R., Poggi, J.-M., Tuleau-Malot, C.: Variable selection using Random Forests. Pattern Recogn. Lett. **31**(14), 2225–2236 (2010)
6. Dash, M., Liu, H.: Feature selection for classification. Intell. Data Analysis **1**(1–4), 131–156 (1997)
7. Das, S.K., Das, S.R.: Filters, wrappers and a boosting-based hybrid for feature selection. In: International Conference on Machine Learning (2001)
8. Dalton, A., ÓLaighin, G.: Comparing supervised learning techniques on the task of physical activity recognition. IEEE J. Biomed. Health Inf. **17**(1), 46–52 (2013)
9. Thabtah, F.: Machine learning in autistic spectrum disorder behavioral research: a review and ways forward. Inform. Health Soc. Care **44**(3), 278–297 (2019)
10. Khan, A.M., Young-Koo Lee, Lee, S.Y., Kim, T.-S.: A triaxial accelerometer-based physical-activity recognition via augmented-signal features and a hierarchical recognizer. IEEE Trans. Inf. Technol. Biomed. **14**(5), 1166–1172 (2010)

11. Darst, B.F., Malecki, K.C., Engelman, C.D.: Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. BMC Genet. **19**(S1) (2018). https://doi.org/10.1186/s12863-018-0633-8
12. Duan, K.B., et al.: Multiple SVM-RFE for gene selection in cancer classification with expression data. IEEE Trans. Nanobiosci. **4**(3), 228–234 (2005)
13. Granitto, P.M., Furlanello, C., Biasioli, F., Gasperi, F.: Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. Chemometr. Intell. Lab. Syst. **83**(2), 83–90 (2006)
14. Guyon, S., et al.: Gene selection for cancer classification using support vector machines. Mach. Learn. **46**(1–3), 389–422 (2002)
15. Fleuret, F.: Binary feature selection with conditional mutual information. Ph.D. thesis. INRIA (2003)
16. Meyer, P.E., Schretter, C., Bontempi, G.: Information-theoretic feature selection in microarray data using variable complementarity. IEEE J. Sel. Top. Sig. Process. **2**(3), 261–274 (2008)
17. Kononenko, I.: Estimating attributes: analysis and extensions of RELIEF. In: Bergadano, F., De Raedt, L. (eds.) ECML 1994. LNCS, vol. 784, pp. 171–182. Springer, Heidelberg (1994). https://doi.org/10.1007/3-540-57868-4_57
18. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artif. Intell. **97**(1–2), 273–324 (1997)
19. Yu, L., Liu, H.: Efficient feature selection via analysis of relevance and redundancy. J. Mach. Learn. Res. **5**(12), 1205–1224 (2005)
20. Fleuret, F.: Fast binary feature selection with conditional mutual information. J. Mach. Learn. Res. **5**(3), 1531–1555 (2004)
21. Estevez, P.A., Tesmer, M., Perez, C.A., Zurada, J.M.: Normalized mutual information feature selection. IEEE Trans. Neural Netw. **20**(2), 189–201 (2009)
22. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. **27**(8), 1226–1238 (2005)
23. Lin, D., Tang, X.: Conditional infomax learning: an integrated framework for feature extraction and fusion. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 68–82. Springer, Heidelberg (2006). https://doi.org/10.1007/11744023_6
24. Kwak, N., Choi, C.-H.: Input feature selection for classification problems. IEEE Trans. Neural Netw. **13**(1), 143 (2002)
25. Bontempi, G., Meyer, P.E.: Causal filter selection in microarray data. In: International Conference on Machine Learning (2010)
26. El Akadi, A., El Ouardighi, A., Aboutajdine, D.: A powerful feature selection approach based on mutual information. Int. J. Comput. Sci. Netw. Secur. **8**(4), 116 (2008)
27. Song, Q., Ni, J., Wang, G.: A fast clustering-based feature subset selection algorithm for high-dimensional data. IEEE Trans. Knowl. Data Eng. **25**(1), 1–14 (2013)
28. Yang, I., Hua, H., Moody, J.: Data visualization and feature selection: new algorithms for nongaussian data. Adv. Neural. Inf. Process. Syst. **12**, 687–693 (2000)
29. Chernbumroong, S., Shuang, C., Yu, H.: Maximum relevancy maximum complementary feature selection for multi-sensor activity recognition. Exp. Syst. Appl. **42**(1), 573–583 (2015)
30. Biau, G.: Analysis of a random forests model. J. Mach. Learn. Res. **13**, 1063–1095 (2012)