A Framework for Feature Selection to Exploit Feature Group Structures

Kushani Perera[1(✉)], Jeffrey Chan[2], and Shanika Karunasekera[1]

[1] University of Melbourne, Melbourne, VIC 3010, Australia
bperera@student.unimelb.edu.au, karus@unimelb.edu.au
[2] RMIT University, Melbourne, VIC 3000, Australia
jeffrey.chan@rmit.edu.au

Abstract. Filter feature selection methods play an important role in machine learning tasks when low computational costs, classifier independence or simplicity is important. Existing filter methods predominantly focus only on the input data and do not take advantage of the external sources of correlations within feature groups to improve the classification accuracy. We propose a framework which facilitates supervised filter feature selection methods to exploit feature group information from external sources of knowledge and use this framework to incorporate feature group information into minimum Redundancy Maximum Relevance (mRMR) algorithm, resulting in *GroupMRMR* algorithm. We show that *GroupMRMR* achieves high accuracy gains over mRMR (up to ∼35%) and other popular filter methods (up to ∼50%). *GroupMRMR* has same computational complexity as that of mRMR, therefore, does not incur additional computational costs. Proposed method has many real world applications, particularly the ones that use genomic, text and image data whose features demonstrate strong group structures.

Keywords: Filter feature selection · Feature groups · Squared $L_{0,2}$ norm minimisation

1 Introduction

Feature selection is proven to be an effective method in preparing high dimensional data for machine learning tasks such as classification. The benefits of feature selection include increasing the prediction accuracy, reducing the computational costs and producing more comprehensible data and models. Among the three main feature selection methods, filter methods are preferred to wrapper and embedded methods in applications where the computational efficiency, classifier independence, simplicity, ease of use and the stability of the results are required. Therefore, filter feature selection remains an interesting topic in many recent research areas such as biomarker identification for cancer prediction and drugs discovery, text classification and predicting defective software [3–5,10,11,16,18] and has growing interest in big data applications [19]; according to the Google

H. W. Lauw et al. (Eds.): PAKDD 2020, LNAI 12084, pp. 792–804, 2020.
https://doi.org/10.1007/978-3-030-47426-3_61

Scholar search results, the number of research papers published related to filter methods in year 2018 is ∼1,800 of which ∼170 are in gene selection area.

Most of the existing filter methods perform feature selection based on the instance-feature data alone [7]. However, in real world datasets, there are external sources of correlations within feature groups which can improve the usefulness of feature selection. For example, the genes in genomic data can be grouped based on the Gene Ontology terms they are annotated with [2] to improve bio-marker identification for the tasks such as disease prediction and drugs discovery. The words in documents can be grouped according to their semantics to select more significant words which are useful in document analysis [14]. The nearby pixels in images can be grouped together based on their spatial locality to improve selection of pixels for image classification. In software data, software metrics can be grouped according to their granularity in the code to improve the prediction of defective software [11,18]. In Sect. 4, using a text dataset as a concrete example, we demonstrate the importance of feature group information for filter feature selection to achieve good classification accuracy.

Although feature group information have been used to improve feature selection in wrapper and embedded approaches [8,12], group information is only rarely used to improve the feature selection accuracy in filter methods. Yu et al. [19] proposes a group based filter method, GroupSAOLA (GSAOLA), yet being an online method, it achieves poor accuracy, which we show experimentally. The common method used by embedded methods to exploit feature group information is minimising the L_1 and L_2 norms of the feature weight matrix, while minimising the classification error. Depending on whether the features are encouraged from the same group [8] or different groups [12], L_1 norm is used to cause inter group or intra group sparsity. Selecting features from different groups is shown to be more effective than selecting features from the same group [12].

Motivated by these approaches, we show that squared $L_{0,2}$ norm minimization of the feature weight matrix can be used to encourage features from different feature groups in filter feature selection. We propose a generic framework which combines existing filter feature ranking methods with feature weight matrix norm minimisation and use this framework to incorporate feature group information in to mRMR objective [7] because mRMR algorithm achieves high accuracy and efficiency at the same time, compared to other filter methods [3,4]. However, the proposed framework can be used to improve any other filter method, such as information gain based methods. As L_0 norm minimization is an NP-hard problem, we propose a greedy feature selection algorithm, *GroupMRMR*, to achieve the feature selection objective, which *has the same computational complexity as the mRMR algorithm. We experimentally show that for the datasets with feature group structures, GroupMRMR obtains significantly higher classification accuracy than the existing filter methods.* Our main contributions are as follows.

- We propose a framework which supports the filter feature selection methods to utilise feature group information to improve their classification accuracy.
- Using the proposed framework, we integrate feature group information into mRMR algorithm and propose a novel feature selection algorithm.

– Through extensive experiments we show that our algorithm obtains significantly higher classification accuracy than the mRMR and existing filter feature selection algorithms for no additional computational costs.

2 Related Work

Utilization of feature group information to improve prediction accuracy has been popular in embedded feature selection [8,12,17]. Among them, algorithms such as GroupLasso [8] encourage features from the same group while algorithms such as Uncorrelated GroupLasso [12] encourage features from different groups. We select the second approach as it is proven to be more effective for real data [12]. Filter feature selection is preferred over wrapper and embedded methods due to their classifier independence, computational efficiency and simplicity, yet have comparatively low prediction accuracy. However, most filter methods select the features based on the instance-feature data alone, which are coded in the data matrix, using information theoretic measures [7,13,15]. Some methods [20] use the feature group concept, yet the groups are also formed using instance-feature data to reduce feature redundancy. None of these methods take advantage of the external sources of knowledge about feature group structures. GSAOLA [19] is an online filter method which exploits feature groups, however we experimentally show that our method significantly outperforms it in terms of accuracy.

3 Preliminaries

In this section and Table 1, we introduce the terms used later in the paper. Let C be the class variable of a dataset, D, and f_i, f_j any two feature variables.

Definition 1. *Given that X and Y are two feature variables in D, with feature values x and y respectively, mutual information between X and Y, is given by* $I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) log \frac{p(x,y)}{p(x)p(y)}$.

Definition 2. *The relevancy of f_i = $Rel(f_i) = I(f_i;C)$.*

Definition 3. *The redundancy between f_i and f_j = $Red(f_i, f_j) = I(f_i; f_j)$.*

Given that $W \in \mathbb{R}^{M \times N}$, W_i is the i^{th} row of W, W_{ij} is the j^{th} element in W_i, the squared $L_{0,2}$ norm of W is defined as $\|W\|_{0,2}^2 = \sum_{i=1}^{M} (\|W_i\|_0)^2 = \sum_{i=1}^{M} N_i^2$ where $N_i = \|W_i\|_0 = \#\ (j|W_{ij} \neq 0)$. For the scenarios in which the rows of W have different importance levels, we define $\|W\|_{0,2}^2 = \sum_{i=1}^{M} \epsilon_i (\|W_i\|_0)^2 = \sum_{i=1}^{M} N_i^2 \epsilon_i$. ϵ_i is the weight of W_i. k is the required number of features.

Table 1. Frequently used definitions

F	Set of all features	I	Set of all feature group indices
S	Selected feature subset, $S \subseteq F$	G_i	Set of features in i^{th} feature group
G	Set of all feature groups	α_i	The weight of the i^{th} feature group

4 Motivation and Background

Ignoring the external sources of correlations within feature groups may result in poor classification accuracy for the datasets whose features show a group behaviour. We demonstrate this using mRMR algorithm as a concrete example, a filter method which otherwise achieves good accuracy.

mRMR Algorithm: mRMR objective for selecting a feature subset $S \subseteq F$ of size k is as follows.

$$\max_{S} \sum_{f \in S} Rel(f) - \frac{1}{|S|} \sum_{f_i, f_j \in S} Red(f_i, f_j) \text{ subject to } |S| = k, k \in \mathbb{Z}^+ \quad (1)$$

To achieve the above objective, mRMR *selects one feature at a time* to maximise the relevancy of the new feature x with the class variable and to minimise its redundancy with the already selected feature set, as shown in Eq. (2).

$$\max_{x} Rel(x) - \frac{1}{|S|} \sum_{f \in S} Red(x, f) \quad (2)$$

Example 1: Consider selecting two features from the dataset in Fig. 1. In this dataset, each document is classified into one of the four types: Botany, Zoology, Physics or Agriculture. The rows represent the feature vector, the words which have occurred in the documents. 1 means the word has occurred within the document (or has occurred with high frequency) and 0 means otherwise.

The relevancies of the features, Apple, Rice, Cow and Sheep are 0.549, 0.443, 0.311 and 0.311, respectively. mRMR first selects **Apple**, which has the highest relevancy. The redundancies of Rice, Cow and Sheep with respect to Apple are 0.07, 0.017 and 0.016, respectively. Therefore, mRMR next selects **Rice**, the feature with the highest relevancy redundancy difference, 0.373 (0.443 - 0.07). Global mRMR optimisation approaches [15] also select {Apple, Rice}.

	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}	d_{11}	d_{12}	d_{13}	d_{14}	d_{15}	d_{16}
Apple	**1**	**1**	**1**	0	0	0	0	0	0	0	0	0	**1**	**1**	**1**	0
Rice	0	**1**	**1**	**1**	0	0	0	0	0	0	0	0	0	0	**1**	**1**
Cow	0	0	0	0	**1**	**1**	0	0	0	0	0	0	0	0	**1**	**1**
Sheep	0	0	0	0	0	0	**1**	**1**	0	0	0	0	**1**	**1**	0	0
Class	**B**	**B**	**B**	**B**	**Z**	**Z**	**Z**	**Z**	**P**	**P**	**P**	**P**	**A**	**A**	**A**	**A**

Fig. 1. Example text document dataset. Column (d_i): a document/instance, Row: a word/feature, Class: document type, 1/0: Occurrence of a word, B: Botany, Z: Zoology, P: Physics, A: Agriculture

mRMR Algorithm {Apple, Rice}						Different Feature Groups {Apple, Sheep}					
Pattern	*B*	*Z*	*P*	*A*	**Class**	**Pattern**	*B*	*Z*	*P*	*A*	**Class**
(a=1, r=0)	25%	0%	0%	50%	**A**	(a=1, s=0)	75%	0%	0%	25%	**B**
(a=0, r=1)	25%	0%	0%	25%	**A, B**	(a=0, s=1)	0%	50%	0%	0%	**Z**
(a=0, r=0)	0%	100%	100%	0%	**P, Z**	(a=0, s=0)	25%	50%	100%	25%	**P**
(a=1, r=1)	50%	0%	0%	25%	**B**	(a=1, s=1)	0%	0%	0%	50%	**A**

Fig. 2. Value pattern probabilities created by different feature subsets in each class, A: Agriculture, B: Botany, P: Physics, Z: Zoology, Class: The class assigned to the value pattern, %: $\frac{\#(\text{x,y})\text{value patterns in class } c}{\#\text{instances in class } c} \times 100$; x, $y \in \{0,1\}$, a: Apple, r: Rice, s: Sheep

Exploiting Feature Group Semantics: Figure 2 shows the value pattern distribution of {Apple, Sheep} and {Apple, Rice} pairs within each class. In {Apple, Sheep}, the highest probability value pattern in each class is different from one another. Therefore, each value pattern is associated with a different class, which helps distinguishing all the document types from one another. In {Apple, Rice}, there is no such distinctive relationship between the value patterns and classes. Using the value pattern distribution, the classification algorithm cannot distinguish between the Zoology and Physics documents and between Agriculture and Botany documents. *This shows that features from different groups have achieved better class discrimination.*

The reason behind the suboptimal result of the mRMR algorithm is its ignorance about the high level feature group structures. The words Apple and Rice form a group as they are plant names. Cow and Sheep form another group as they are animal names. The documents are classified according to whether they contain plant names or/and animal names, regardless of the exact plant or animal name they contain. Botany documents (d_1–d_4) contain plant names (Apple or Rice) and no animal names. Zoology documents (d_5–d_8) contain animal names (Cow or Sheep) and no plant names. This high level insight is not captured by the instance-feature data alone. Using feature group information as an external source of knowledge and encouraging features from different feature groups help solving this problem.

5 Proposed Method: GroupMRMR

We propose a framework which facilitates filter feature selection methods to exploit feature group information to achieve better classification accuracy. Using this framework, we extend mRMR algorithm into *GroupMRMR* algorithm, which encourages features from different groups to bring in different semantics which help selecting a more balanced set of features. We select mRMR algorithm for extension because it has proven good classification accuracy with low computation costs, compared to other filter feature selection methods. The feature groups are assigned weights (α_i) to represent their importance levels, and *GroupMRMR* selects more features from the groups with higher importance. Group weights may be decided according to factors such as group size and group quality. For this

paper, we assume that the feature groups do not overlap but plan to investigate overlapping groups in the future.

5.1 Feature Selection Objective

Our feature selection objective includes both the filter feature selection objective and encouraging features from different feature groups. To encourage features from different groups, we minimise $\|W\|_{0,2}^2$ of the feature weight matrix, W. Using L_0 norm at intra group level enforces intra group sparsity, discouraging features to be selected from the same group. Using L_2 norm at inter group level encourages features from different feature groups [12].

Let $W \in \mathbb{R}^{|G|\times|F|}$ be a feature weight matrix such that $W_{ij} = 1$ if $f_j \in S$ and $f_j \in G_i$. Otherwise, $W_{ij} = 0$. Given that $g(W)$ is any maximisation quantity used in an existing filter feature selection objective which can be expressed a function of W and λ is a user defined parameter, our objective is to select $S \subseteq F$ to maximise the following subject to $|S| = k$, $k \in \mathbb{Z}^+$:

$$\max_S h(S) = g(W) - \lambda\|W\|_{0,2}^2 \tag{3}$$

Given that $R1 \in \mathbb{R}^{|F|\times|F|}$ is a diagonal matrix in which $R1_{jj} = Rel(f_j)$ and $R2 \in \mathbb{R}^{|F|\times|F|}$ such that $R2_{ij} = Red(f_i, f_j)$ for $i \neq j$ $R1_{ij} = 0$ for $i = j$, it can be shown that $\|WR1W^T\|_{1,1}$ - $\frac{1}{2|S|}\|WR2W^T\|_{1,1} = \sum_{f\in S} Rel(f)$ - $\frac{1}{|S|}\sum_{f_i,f_j\in S} Red(f_i, f_j)$, where W^T is the transpose of W. That is, the maximisation quantity in mRMR objective in Eq. (1) is a function of W. Consequently, $g(W)$ in Eq. (3) can be replaced with the mRMR objective as shown in Eq. (4).

$$\max_S h(S) = \sum_{f\in S} Rel(f) - \frac{1}{|S|}\sum_{f_i,f_j\in S} Red(f_i, f_j) - \lambda\|W\|_{0,2}^2 \tag{4}$$

Definition 4. *Given that S and G_i are as defined in Table 1, $n_i = |S \cap G_i| =$ No. of features in S and G_i.*

Given n_i is as defined in Definition 4, according to Sect. 3, $\|W\|_{0,2}^2 = \sum_{i=1}^{|G|} n_i^2$. When the feature groups have different weights, the rows of W also have different importance levels. In such scenarios, $\|W\|_{0,2}^2 = \sum_{i=1}^{|G|} n_i^2\epsilon_i$, where $\epsilon_i = \frac{1}{\alpha_i}$ where $\alpha_i > 0$. Consequently, we can rewrite the objective in Eq. (4) as in Eq. (5) subject to $|S| = k$, $k \in \mathbb{Z}^+$. As the feature groups do not overlap, $\sum_{i=1}^{|G|} n_i = |S|$. Using Eq. (5), we present Theorem 1 that shows minimising $\|W\|_{0,2}^2$ is equivalent to encouraging features from different groups in to S.

$$\max_S h(S) = \sum_{f\in S} Rel(f) - \frac{1}{|S|}\sum_{f_i,f_j\in S} Red(f_i, f_j) - \lambda\sum_{i=1}^{|G|}\frac{n_i^2}{\alpha_i} \tag{5}$$

Theorem 1. *Given $\sum_{i=1}^{|G|} n_i = |S| = k$, minimum $\sum_{i=1}^{|G|} \frac{n_i^2}{\alpha_i}$ is obtained when $\frac{n_i}{\alpha_i} = \frac{n_j}{\alpha_j}$, $\forall\, i, j \in I$, where $k \in \mathbb{Z}^+$ is a constant.*

Algorithm 1. GroupMRMR algorithm

input : Dataset (D), Required feature count (r), Group weights ($\alpha_1 \cdots \alpha_{|G|}$)
output: Selected feature subset (S)

$U \leftarrow F$ in D; $feaCount \leftarrow 0$; $n_1 \cdots n_{|G|} \leftarrow 0$;
while $feaCount < r$ **do**
  for $x \in U$ **do**
    $p \leftarrow$ Group index of G_p where $x \in G_p$;
    $score_x \leftarrow Rel(x)$ - $\frac{1}{|S|} \sum_{f \in S} Red(x; f)$ - $\lambda \frac{2n_p+1}{\alpha_p}$;
  end
  $f_{max} \leftarrow \text{argmax}_{x \in U}\ score_x$;
  $S \leftarrow S + f_{max}$; $U \leftarrow U$ - f_{max};
  $j \leftarrow$ Group index of G_j where $f_{max} \in G_j$;
  n_j++; $feaCount$++;
end
return S;

Proof. Using Lagrange multipliers method, we show minimum $\sum_{i=1}^{|G|} \frac{n_i^2}{\alpha_i}$ is achieved when $\frac{n_1}{\alpha_1} = \frac{n_2}{\alpha_2} = \cdots = \frac{n_{|G|}}{\alpha_{|G|}}$. Please refer to this link[1] for the detailed proof.

5.2 Iterative Feature Selection

As $L^2_{0,2}$ minimisation is NP-hard, we propose a heuristic algorithm to achieve the objective in Eq. (4). The algorithm selects a feature, f_t, at each iteration t to maximise the difference between $h(S_t)$ and $h(S_{t-1})$, where S_t and S_{t-1} are the feature subsets selected after Iteration t and $t-1$ respectively and $h(.)$ is as defined in Eq. (5). As there are datasets with millions of features *we seek an algorithm to select f_t with linear complexity.* Theorem 2 shows that $h(S_t)$ - $h(S_{t-1})$ can be maximised by adding the term, $\lambda \frac{2n_p+1}{\alpha_p}$ to the mRMR algorithm in Eq. (2). p is the feature group of the evaluated feature (f_x), n_p is the number of features already selected from p before Iteration t and α_p is the weight of p.

Theorem 2. *Given that S_t, S_{t-1}, $h(S_t)$, $h(S_{t-1})$, p, n_p, α_p as defined above and S'_{t-1} is the unselected feature subset after Iteration $t-1$,* $\text{argmax}_{f_{x \in S'_{t-1}}}$ $h(S_t)$ - $h(S_{t-1})$ = $\text{argmax}_{f_{x \in S'_{t-1}}}$ $Rel(f_x; c)$ - $\frac{1}{|S_{t-1}|} \sum_{f_i \in S_{t-1}} Red(f_x; f_i)$ - $\lambda \left(\frac{2n_p+1}{\alpha_p}\right)$.

Proof. To prove this, we use the fact that $|S_t|$ and $|S_{t-1}|$ are constants at a given iteration. Please refer to this link (see footnote 1) for the detailed proof.

[1] https://sites.google.com/view/kushani/publications.

Table 2. Dataset description. m: # features, n: # instances, c: # classes

Dataset	m	n	c	Type	Dataset	m	n	c	Type
Multi-Tissue (MT) [1]	1,000	103	4	Genomic	CNS [1]	989	42	5	Genomic
Leukemia (LK) [1]	999	38	3	Genomic	Yale [6]	1,024	165	15	Image
Multi-A [1]	5,565	103	4	Genomic	BBC [9]	9,635	2,225	5	Text
Groovy (GRV) [18]	65	757	2	Software					

Based on Theorem 2, we propose *GroupMRMR* algorithm. At each iteration, the feature score of each feature in U is computed as shown in Line 5 of Algorithm 1. The feature with the highest score is removed from U and added to S (Line 7–10 in Algorithm 1). The algorithm can be modified to encourage the features from the same group as well by setting $\lambda < 0$.

Example 1 Revisited: Next, we apply *GroupMRMR* for Example 1. We assume $\lambda = 1$ and $\alpha_i = \alpha_j = 1, \forall\ i, j \in I$. *GroupMRMR* first selects Apple, the feature with highest relevancy (0.549). In Iteration 2, n_p value for Rice, Cow, and Sheep are 1, 0 and 0, respectively and $\frac{2n_p+1}{\alpha_p}$ are 3, 0 and 0, respectively. The redundancies of each feature with Apple are same as computed in Sect. 4. The feature scores for Rice, Cow and Sheep are -2.627 (0.443-0.07-3), 0.294 (0.311-0.017-0) and 0.295 (0.311-0.016-0), respectively and *GroupMRMR* selects Sheep, the feature with the highest feature score. Therefore, *GroupMRMR* selects {Apple, Sheep}, the optimal feature subset, as discussed in Sect. 4.

Computation Complexity: The computational complexity of *GroupMRMR* is the same as that of mRMR, which is $O(|S||F|)$. $|S|$ and $|F|$ are the cardinalities of the selected feature subset and the complete feature set, respectively. As $|S| << |F|$, *GroupMRMR* is effectively linear with $|F|$.

6 Experiments

This section discusses the experimental results for *GroupMRMR* for real datasets.

Datasets: We evaluate *GroupMRMR*, using real datasets, which are benchmark datasets used to test group based feature selection. Table 2 shows a summary of them. Images in Yale have a 32×32 pixel map. GRV is a JIRA software defect dataset whose features are code quality metrics.

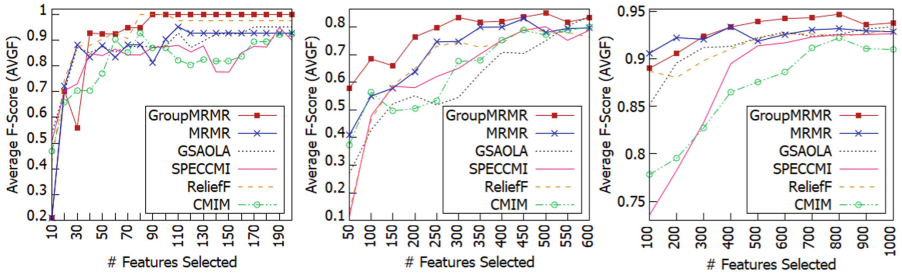
Grouping Features: The pixel map of the images are partitioned into $m \times m$ non overlapping squares such that each square is a feature group. This introduces spatial locality information, not available from just the data (instance-feature)

Table 3. Comparison of accuracies achieved by different algorithms. Row 1: The maximum accuracy (in AVGF) gained by each algorithm in each dataset. The highest maximum AVGF for each dataset is in bold letters. Row 2 (x): the number of features at which the highest AVGF is achieved. Row 3 (%): The average accuracy gain of *GroupMRMR* over the baseline. +: *GroupMRMR* wins, −: *GroupMRMR* losses

	MT	CNS	LK	Multi-A	Yale	BBC	GRV
GroupMRMR	**1**	**0.9**	**1**	**1**	**0.85**	**0.95**	**0.66**
	(110)	(90)	(20)	(90)	(500)	(800)	(10)
MRMR	0.98	0.88	0.94	0.95	0.83	0.93	0.57
	(70)	(180)	(40)	(110)	(450)	(400)	(30)
	+4%	+11%	+4%	+5%	+7%	0%	+4%
GSAOLA	0.95	0.86	1	0.95	0.84	0.93	0.56
	(60)	(50)	(50)	(170)	(600)	(1000)	(25)
	+1%	+2%	+2%	+3%	+17%	+3%	+3%
SPECCMI	0.9	0.71	1	0.95	0.80	0.93	0.61
	(90)	(180)	(190)	(190)	(500)	(1000)	(30)
	+12%	+16%	+17%	+8%	+14%	+7%	−1%
CMIM	0.95	0.83	0.88	0.93	0.8	0.92	0.61
	(200)	(160)	(90)	(80)	(600)	(800)	(25)
	+10%	+19%	+32%	+9%	+13%	+8%	−1%
ReliefF	0.95	0.83	1	**1**	0.8	0.93	0.52
	(60)	(170)	(80)	(80)	(450)	(1000)	(25)
	+2%	+6%	+3%	−1%	+12%	+2%	+6%

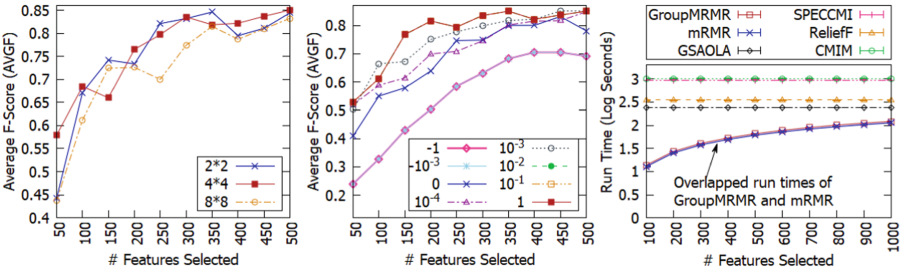
itself. The genes in genomic data are clustered based on the Gene Ontology term annotations as described in [2]. The number of groups is set to 0.04 of the original feature set, based on the previous findings for MT dataset [2]. Words in BBC dataset are clustered using k-means algorithm, based on the semantics available from Word2Vec [14]. We use only 2,411 features, only the words available in the Brown's corpus. Number of word groups is 50, which is selected by cross validation results on the training data. The code metrics in software defect data are grouped into five groups based on their granularity in the code [18].

Baselines: We compare *GroupMRMR* with existing filter methods which have proven high accuracy. mRMR algorithm, of which the *GroupMRMR* is an extension, is a greedy approach to achieve mRMR objective while SPECCMI [15] is a global optimisation algorithm to achieve the same. Conditional Mutual Information (CMIM) [15] is a mutual information based filter method not belonging to the mRMR family. ReliefF [13] is a distance based filter method. GSAOLA [19] is an online filter method which utilises feature group information.



(a) Multi-A (b) Yale (c) BBC

Fig. 3. Classification accuracy variation with the number of selected features



(a) Group size (b) λ parameter (c) Run time

Fig. 4. Accuracy and runtime variations for Yale and BBC datasets (a) Accuracy variation with the group size (Yale) (b) Accuracy variation with λ (Yale) (c) Average run time variation (in log scale) of the algorithms (BBC). 95% confidence interval error bars are too small to be visible due to the high precision (standard deviations $\sim$2 s)

Evaluation Method: The classifier's prediction accuracy on the test dataset with selected features is considered as the prediction accuracy of the feature selection algorithm. It is measured in terms of the Macro-F1, the average of the F1-scores for each class (AVGF). Average accuracy is the average of AVGFs for all the selected feature numbers up to the point algorithm accuracies converge. The log value of the average run time (measured in seconds) is reported.

Experimental Setup: We split each dataset, 60% instances for training set and 40% for test set, using stratified random sampling method. Feature selection is performed on the training set and the classifier is trained on the training set with the selected features. The classifier is then used to predict the labels of the test set. Due to the small sample size of the datasets we do not use a separate validation set for tuning λ. Instead, we select $\lambda \in [0, 2]$, which gives the highest classification accuracy on the training set. The classifier used is the Support Vector Machine. For image data, default $m = 4$. For genomic data, $\alpha_i = 1, \forall\ i$. For other datasets, $\alpha_i = \frac{|G_i|}{|F|}$ (G_i,F are defined in Table 1).

Experiment 1: Measures the classification accuracy obtained for the datasets with selected features. ***Experiment 2:*** Performs feature selection for image datasets with different feature group sizes: $m \times m$ (m = 2,4,8). This tests the *effect of the group size* on the classification accuracy. ***Experiment 3:*** Runs *GroupMRMR* for different $\lambda \in [-1, 1]$. This tests the *effect of* λ on the classification accuracy. ***Experiment 4:*** Executes each feature selection algorithm 20 times and compute the average run time to evaluate *algorithm efficiency.*

Experimental Results: Table 3 shows that *GroupMRMR achieves the highest AVGF in all datasets over baselines.* In LK dataset, the 100% accuracy is achieved with a lower number of features than baselines. *GroupMRMR* achieves higher or same average accuracy compared to baselines in 32 out of 35 cases. Figure 3 shows that, despite the slightly low average accuracy compared to ReliefF, *GroupMRMR* maintains a higher accuracy than baselines in Multi-A for most of the selected feature numbers. Other datasets also show similar results, yet we show only three graphs due to the space limitations. Please refer to this link (see footnote 1) to see all the results graphs. The maximum accuracy gain of *GroupMRMR* over the accuracy gained by the complete feature set is 2%, 10%, 2%, 2%, 1% and 6% for MT, CNS, Multi-A, Yale, BBC and GRV datasets, respectively. The maximum accuracy gain of *GroupMRMR* is 50% over SPECCMI in Yale dataset at 50 selected features. The highest accuracy gain of *GroupMRMR* over mRMR is 35% in CNS dataset at 70 selected features. Figure 4a shows that the classification accuracy of *GroupMRMR* for 8×8 image partitions is less than for 4×4 and 2×2 partitions. Figure 4b shows that the *classification accuracy is not much sensitive to* λ in the $[10^{-3}, 1]$ range, yet degrades to a large extent when $\lambda < 0$. Figure 4c shows that the *runtime of GroupMRMR is almost the same as the run time of mRMR algorithm and lower than most of the other baseline methods* (∼10 times lower than SPECCMI and CMIM for BBC dataset).

Evaluation Insights: *GroupMRMR consistently shows good classification accuracy compared to baselines* for all the datasets (highest average accuracy and highest maximum accuracy in almost all datasets). The equal run times of *GroupMRMR* and mRMR show that *the accuracy gain is obtained for no additional costs* and supports the time complexity analysis in Sect. 5. Better prediction accuracy is obtained for small groups because large feature groups resemble the original feature set with no groupings. This shows the importance of feature group information to gain high feature selection accuracy. The *accuracy is lower when the features are encouraged from the same group* ($\lambda < 0$) *instead from different groups* ($\lambda > 0$), which supports our hypothesis. The classification accuracy is less sensitive to $\lambda \geq 10^{-3}$, therefore *parameter tuning is less required.*

7 Conclusion

We propose a framework which facilitates filter feature selection methods to exploit feature group information as an external source of information. Using this framework, we incorporate feature group information into mRMR algorithm, resulting in *GroupMRMR* algorithm. We show that compared to baselines, *GroupMRMR* achieves high classification accuracy for the datasets with feature group structures. The run time of *GroupMRMR* is same as the run time of mRMR, which is lower than many existing feature selection algorithms. Our future work include experimenting the proposed framework for other filter methods and detecting whether a dataset contains feature group structures.

Acknowledgements. This work is supported by the Australian Government.

References

1. Cancer program datasets. http://portals.broadinstitute.org/cgi-bin/cancer/datasets.cgi. Accessed Nov 2019
2. Acharya, S., Saha, S., Nikhil, N.: Unsupervised gene selection using biological knowledge: application in sample clustering. BMC Bioinform. **18**(1), 513 (2017)
3. Alirezanejad, M., Enayatifar, R., Motameni, H., et al.: Heuristic filter feature selection methods for medical datasets. Genomics (2019). https://doi.org/10.1016/j.ygeno.2019.07.002
4. Bolón-Canedo, V., Rego-Fernández, D., Peteiro-Barral, D., Alonso-Betanzos, A., Guijarro-Berdiñas, B., Sánchez-Maroño, N.: On the scalability of feature selection methods on high-dimensional data. Knowl. Inf. Syst. **56**(2), 395–442 (2017). https://doi.org/10.1007/s10115-017-1140-3
5. Bommert, A., Sun, X., Bischl, B., et al.: Benchmark for filter methods for feature selection in high-dimensional classification data. CSDA **143**, 106839 (2020)
6. Cai, D., He, X., Hu, Y., et al.: Learning a spatially smooth subspace for face recognition. In: Proceedings of IEEE CVPR 2007, pp. 1–7 (2007)
7. Ding, C., Peng, H.: Minimum redundancy feature selection from microarray gene expression data. JBCB **3**(02), 185–205 (2005)
8. Friedman, J., Hastie, T., Tibshirani, R.: A note on the group lasso and a sparse group lasso. arXiv preprint arXiv:1001.0736 (2010)
9. Greene, D., Cunningham, P.: Practical solutions to the problem of diagonal dominance in kernel document clustering. In: Proceedings of the 23rd ICML, pp. 377–384 (2006). https://doi.org/10.1145/1143844.1143892
10. Hancer, E., Xue, B., Zhang, M.: Differential evolution for filter feature selection based on information theory and feature ranking. Knowl.-Based Syst. **140**, 103–119 (2018). https://doi.org/10.1016/j.knosys.2017.10.028
11. Jiarpakdee, J., Tantithamthavorn, C., Treude, C.: Autospearman: Automatically mitigating correlated metrics for interpreting defect models. arXiv preprint arXiv:1806.09791 (2018)
12. Kong, D., Liu, J., Liu, B., et al.: Uncorrelated group lasso. In: AAAI, pp. 1765–1771 (2016)
13. 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

14. Lilleberg, J., Zhu, Y., Zhang, Y.: Support vector machines and word2vec for text classification with semantic features. In: Proceedings of the 14th IEEE ICCI* CC, pp. 136–140 (2015). https://doi.org/10.1109/ICCI-CC.2015.7259377
15. Nguyen, X.V., Chan, J., Romano, S., et al.: Effective global approaches for mutual information based feature selection. In: Proceedings of the 20th ACM SIGKDD, pp. 512–521 (2014). https://doi.org/10.1145/2623330.2623611
16. Uysal, A.K., Gunal, S.: A novel probabilistic feature selection method for text classification. Knowl.-Based Syst. **36**, 226–235 (2012)
17. Wang, J., Wang, M., Li, P., et al.: Online feature selection with group structure analysis. IEEE TKDE **27**(11), 3029–3041 (2015)
18. Yatish, S., Jiarpakdee, J., Thongtanunam, P., et al.: Mining software defects: should we consider affected releases? In: Proceedings of the 41st International Conference on Software Engineering, pp. 654–665. IEEE Press (2019)
19. Yu, K., Wu, X., Ding, W., et al.: Scalable and accurate online feature selection for big data. ACM TKDD **11**(2), 16 (2016). https://doi.org/10.1145/2976744
20. Yu, L., Ding, C., Loscalzo, S.: Stable feature selection via dense feature groups. In: Proceedings of the 14th ACM SIGKDD, pp. 803–811 (2008). https://doi.org/10.1145/1401890.1401986