A New Fuzzy-Rough Hybrid Merit to Feature Selection

Javad Rahimipour Anaraki[1](✉), Saeed Samet[2], Wolfgang Banzhaf[3], and Mahdi Eftekhari[4]

[1] Department of Computer Science, Memorial University of Newfoundland,
St. John's, Nl A1B 3X5, Canada
jra066@mun.ca

[2] Faculty of Medicine, Memorial University of Newfoundland,
St. John's, Nl A1B 3V6, Canada
saeed.samet@med.mun.ca

[3] Department of Computer Science, Memorial University of Newfoundland,
St. John's, Nl A1B 3X5, Canada
banzhaf@mun.ca

[4] Department of Computer Engineering, Shahid Bahonar University of Kerman,
7616914111 Kerman, Iran
m.eftekhari@uk.ac.ir

Abstract. Feature selecting is considered as one of the most important pre-process methods in machine learning, data mining and bioinformatics. By applying pre-process techniques, we can defy the curse of dimensionality by reducing computational and storage costs, facilitate data understanding and visualization, and diminish training and testing times, leading to overall performance improvement, especially when dealing with large datasets. Correlation feature selection method uses a conventional merit to evaluate different feature subsets. In this paper, we propose a new merit by adapting and employing of correlation feature selection in conjunction with fuzzy-rough feature selection, to improve the effectiveness and quality of the conventional methods. It also outperforms the newly introduced gradient boosted feature selection, by selecting more relevant and less redundant features. The two-step experimental results show the applicability and efficiency of our proposed method over some well known and mostly used datasets, as well as newly introduced ones, especially from the UCI collection with various sizes from small to large numbers of features and samples.

Keywords: Feature selection · Fuzzy-rough dependency degree · Correlation merit

1 Introduction

Each year the amount of generated data increases dramatically. This expansion needs to be handled to minimize the time and space complexities as well as the

J.F. Peters and A. Skowron (Eds.): TRS XX, LNCS 10020, pp. 1–23, 2016.
DOI: 10.1007/978-3-662-53611-7_1

comprehensibility challenges inherent in big datasets. Machine learning methods tend to sacrifice some accuracy to decrease running time, and to increase the clarity of the results [1].

Datasets may contain hundreds of thousand of samples with thousands of features that make further processing on data a tedious job. Reduction can be done on either features or on samples. However, due to the high cost of sample gathering and their undoubted utility, such as in bioinformatics and health systems, data owners usually prefer to keep only the useful and informative features and remove the rest, by applying Feature Selection (FS) techniques that are usually considered as a preprocessing step to further processing (such as classification). These methods lead to less classification errors or at least to minimal diminishing of performance [2].

In terms of data usability, each dataset contains three types of features: 1- informative, 2- redundant, and 3- irrelevant. Informative features are those that contain enough information on the classification outcome. In other words, they are non-redundant, relevant features. Redundant features contain identical information compared to other features, whereas irrelevant features have no information about the outcome. The ideal goal of FS methods is to remove the last two types of features [1].

FS methods can generally be divided into two main categories [3]. One approach is *wrapper* based, in which a learning algorithm estimates the accuracy of the subset of features. This approach is computationally intensive and slow due to the large number of executions over selected subsets of features, that make it impractical for large datasets. The second approach is *filter* based, in which features are selected based on their quality regardless of the results of learning algorithm. As a result, it is fast but less accurate. Also, a combinational approach of both methods called *embedded* has been proposed to accurately handle big datasets [4]. In the methods based on this approach, feature subset selection is done while classifier structure is being built.

One of the very first feature selection methods for binary classification datasets is Relief [5]. This method constructs and updates a weight vector of a feature, based on the nearest feature vector of the same and different classes using Euclidean distance. After a predefined number of iterations l, relevant vector is calculated by dividing the weight vector by l, and the features with relevancy higher than a specific threshold will be selected. Hall [1] has proposed a merit based on the average intra-correlation of features and inter-correlation of features and the outcome. Those features with higher correlation to the outcome and lower correlation to other features are selected.

Jensen et al. [6] have introduced a novel feature selection method based on lower approximation of the fuzzy-rough set, in which features and outcome dependencies are calculated using a merit called Dependency Degree (DD). In [7], two modifications of the fuzzy-rough feature selection have been introduced to improve the performance of the conventional method: 1- Encompassing the selection process in *equal* situations, where more than one feature result in an identical fitness value by using correlation merit [1] and 2- Combining the first improvement with the stopping criterion [8]. Qian et al. [9], have proposed an accelerator to perform sample and feature selection simultaneously in order to improve the time complexity of fuzzy-rough feature selection. Jensen et al. [10] have designed a new version of fuzzy-rough feature selection to deal with semi-supervised datasets, in which class feature is partially labelled. Shang et al. [11] have introduced a hybrid system for Mars images based on conjunction of fuzzy-rough feature selection and support vector machines. The behaviour of k-nearest neighbours classifier has been improved by Derrac et al. [12], using fuzzy rough feature selection and steady-state genetic algorithm for both feature and prototype selection. Dai et al. [13], have designed a system using fuzzy information gain ratio based on fuzzy rough feature selection structure to classify tumor data in gene expression.

Xu et al. [14] have proposed a non-linear feature selection method based on gradient boosting of limited depth trees. This method combines classification and feature selection processes into one by using gradient boosted regression trees resulting from the greedy CART algorithm.

In this paper, we propose a new merit, which is not only capable of effectively removing redundant features, selecting relevant ones, and enhancing the classification accuracy, but it also outperforms when applied to large datasets, compared to the other existing methods.

In Sect. 2, background and preliminaries of correlation based and fuzzy-rough feature selection methods are described in detail. Our proposed method is discussed in Sect. 3. Section 4 is dedicated to experimental results and discussion on the performance and effectiveness of the new approach comparing with previously introduced methods. Conclusions and future directions are explained in Sect. 5.

2 Preliminaries

In this section, the idea and explanation of the Correlation-based Feature Selection (CFS) method will be presented in Sect. 2.1. Subsection 2.2 illustrates the rough set theory and the rough set based feature selection approach.

2.1 Correlation Based Feature Selection (CFS)

In the feature selection process, selecting those features that are highly correlated with the class attribute while loosely correlated with the rest of the features, is the ultimate goal. One of the most successful feature selection methods based on this is CFS [1]. The evaluation measure of CFS is designed in such a way that it selects predictive and low level inter-correlated features on the class and other features, respectively. Equation 1 shows the merit.

$$Merit_S = \frac{k\overline{r}_{cf}}{\sqrt{k + k(k-1)\overline{r}_{ff}}}, \tag{1}$$

where S is a subset of features, k is the number of selected features, $\overline{r}_{cf}$ is the mean of the correlations of the selected features to the class attribute, and $\overline{r}_{ff}$ is the average of inter-correlations of features. The enumerator calculates how much the selected subset is correlated with the class, and the denominator controls the redundancy of selected features within the subsets. At the heart of the merit, correlation undeniably plays the most important role. Therefore, maximizing merit requires the most relevant features (to maximize the numerator) and the least redundant ones (to minimize the denominator) to be included in the subset. The relevancy and non-redundancy are two important factors in feature selection that are handled in CFS. However, correlation is only capable of measuring linear relationships of two vectors [15]; therefore, in the case of non-linear relationships, the result will be inaccurate.

2.2 Rough Set Feature Selection

The rough set theory has been proposed by Pawlak that is a mathematical tool to handle vagueness in effective way [16]. Suppose U and A to be the universe of discourse and a nonempty set of attributes, respectively, and the information system is presented by $I = (U, A)$. Consider X as a subset of U, and P and Q as subsets of A; approximating a subset in rough set theory is done through the lower and upper approximations. The lower approximation of X, $(\underline{P}X)$ involves those objects which are surely classified in X with regarding to attributes in P. Whereas, upper approximation of X, $(\overline{P}X)$ accommodates those objects which can possibly classified in X considering attributes of P. By defining the lower and upper approximations, a rough set is shown using an ordered pair $(\underline{P}X, \overline{P}X)$. Based on these approximations, different regions in rough set theory is illustrated by Eqs. 2, 3 and 4.

The union of all objects in different regions of $\mathbb{U}$ partitioned by Q with regarding to P is called positive region $POS_P(Q)$.

$$POS_P(Q) = \bigcup_{X \in \mathbb{U}/Q} \underline{P}X \tag{2}$$

The negative region is collection of object that are in $\mathbb{U}$ but not in $POS_P(Q)$, and is shown by $NEG_P(Q)$ [17].

$$NEG_P(Q) = \mathbb{U} - \bigcup_{X \in \mathbb{U}/Q} \overline{P}X \tag{3}$$

The boundary region has determinative role in specifying the type of a set. If the region is a non-empty set, it is called a rough set, otherwise, it is a crisp set.

$$BND_P(Q) = \bigcup_{X \in \mathbb{U}/Q} \overline{P}X - \bigcup_{X \in \mathbb{U}/Q} \underline{P}X \tag{4}$$

The rough set theory can be used to measure the magnitude of dependency between attributes. The dependency of attributes in Q on attribute(s) in P is shown by $P \Rightarrow_k Q$, in which k equals to $\gamma_P(Q)$ and it is labeled Dependency

Degree (DD) [17]. If $0 < k < 1$, then Q partially depends on P, otherwise if $k = 1$ then Q completely depends on P. Equation 5 calculates the DD of Q on P.

$$k = \gamma_P(Q) = \frac{|POS_P(Q)|}{|\mathbb{U}|}, \tag{5}$$

where notation $|.|$ is number of objects in a set.

The reduct set is a subset of features which has identical DD as considering all features. The members of the reduct set are the most informative features which feature outcome is highly dependent on them, while non-members are irrelevant and/or redundant ones.

The most important drawback of rough set based FS methods is their incapability of handing continuous data. One way to govern this imperfection is to discretize continuous data in advance that is necessary but not enough, as long as the amount of similarity between discretized data is unspecified. The ultimate way to handle continuous data using rough set theory is fuzzy-rough set. To begin with, the definition of the X-lower and X-upper approximations and the degree of fuzzy similarity [6] are given by Eqs. 6 to 8, respectively

$$\mu_{\underline{P}X}(x) = \inf_{y \in \mathbb{U}} I\{\mu_{R_P}(x,y), \mu_X(y)\}, \tag{6}$$

$$\mu_{\overline{P}X}(x) = \sup_{y \in \mathbb{U}} T\{\mu_{R_P}(x,y), \mu_X(y)\}, \tag{7}$$

$$\mu_{R_P}(x,y) = \bigcap_{a \in P} \{\mu_{R_a}(x,y)\}, \tag{8}$$

where I is a Łukasiewicz fuzzy *implicator*, which is defined by $min(1-x+y,1)$, and T is a Łukasiewicz fuzzy t-norm, which is defined by $max(x+y-1,0)$. In [18], three classes of fuzzy-rough sets based on three different classes of implicators, namely S-, R-, and QL-implicators, and their properties have been investigated. Here, R_P is the fuzzy similarity relation considering the set of features in P, and $\mu_{R_P}(x,y)$ is the degree of similarity between objects x and y over all features in P. Also, $\mu_X(y)$ is the membership degree of y to X. One of the best fuzzy similarity relations as suggested in [6] is given by Eq. 9.

$$\mu_{R_a}(x,y) = max\left\{min\left\{\frac{(a(y)-(a(x)-\sigma_a))}{\sigma_a}, \frac{((a(x)+\sigma_a)-a(y))}{\sigma_a}\right\}, 0\right\} \tag{9}$$

where σ_a is variance of feature a. Definitions of fuzzy lower and upper approximations are the same as rough lower and upper approximations, except the fact that fuzzy approximations deal with fuzzy values, operators, and output; however, rough approximations deal with discrete and categorical values.

The positive region in the rough set theory is defined as a union of lower approximations. By referring to the extension principle [6], the membership of object x to a fuzzy positive region is given by Eq. 10.

$$\mu_{POS_P(Q)}(x) = \sup_{X \in \mathbb{U}/Q} \mu_{\underline{P}X}(x) \tag{10}$$

where supremum of lower approximations of all partitions resulting from U/Q construct positive region.

If the equivalence class that includes x does not belong to a positive region, clearly x will not be part of a positive region. Using the definition of positive region, the FRDD function is defined as:

$$\gamma'_P(Q) = \frac{|\mu_{POS_P(Q)}(x)|}{|\mathbb{U}|} = \frac{\sum_{x \in \mathbb{U}} \mu_{POS_P(Q)}(x)}{|\mathbb{U}|} \tag{11}$$

where notation $|.|$ is number of objects in a set; however, in numerator we are dealing with fuzzy values and cardinality can be calculated using summation. For denominator $|\mathbb{U}|$ is size of samples in dataset.

The Lower approximation Fuzzy-Rough Feature Selection (L-FRFS) as shown in Algorithm 1 is based on FRDD as shown in Eq. 11, and greedy forward search algorithm, which is capable of being applied to real-valued datasets. The L-FRFS algorithm finds a reduct set without finding all the subsets [6]. It begins with an empty set and each time selects the feature that causes the greatest increase in the FRDD. The algorithm stops when adding more features does not increase the FRDD. Since it employs a greedy algorithm, it does not guarantee that the minimal reduct set will be found. For this reason, a new feature selection merit presented in this section.

Algorithm 1. Lower approximation Fuzzy-Rough Feature Selection

```
$C$, the set of all conditional attributes
$D$, the set of decision attributes
$R \leftarrow \{\}; \gamma'_{best} = 0; \gamma'_{prev} = 0$
do
  $T \leftarrow R$
  $\gamma'_{prev} \leftarrow \gamma'_{best}$
  foreach $x \in (C - R)$
    if $\gamma'_{R \cup \{x\}}(D) > \gamma'_T(D)$
      $T \leftarrow R \cup \{x\}$
      $\gamma'_{best} \leftarrow \gamma'_T(D)$
   $R \leftarrow T$
until $\gamma'_{best} = \gamma'_{prev}$
return $R$
```

3 Proposed Method

On the one hand, FRDD is capable and effective in uncovering the dependency of a feature to another, and the feature selection method based on the merit has shown remarkable performance on resulting classification accuracies [6]. The L-FRFS algorithm evaluates every feature to find the one with the highest dependency, and continues the search by considering every features combination to asset the most dependent features subset to the outcome. However, tracking and finding highly dependent features to the class might end in selecting redundant features.

On the other hand, CFS merit, as shown in Eq. 1, has the potentiality of selecting less redundant features due to the structure of the denominator, in which the square root of the mean of the correlation of the features to each other has a positive impact on the number of redundant features being selected.

By considering capabilities of CFS merit, substituting the correlation with Fuzzy-Rough Dependency Degree (FRDD) that is fuzzy version of DD could take advantage of both criteria to construct a more powerful merit. In this section, the proposed approach is defined based on the two main concepts of feature selection: 1- Evaluation measure, and 2- Search method. The evaluation measure is the new hybrid merit and the search method is hill-climbing.

3.1 A New Hybrid Merit

Based on the concepts of the FRDD and CFS, we have developed a new hybrid merit by substituting the correlation in CFS with FRDD to benefit from both merits. Equation 12 shows the proposed merit.

$$\delta = \frac{\sum_{i=1}^{k} \gamma'_i(c)}{\sqrt{k \times \left(1 + \sum_{j=1}^{k-1} \gamma'_j(f)\right)}}, \tag{12}$$

where $\gamma'_i(c)$ is the FRDD of already selected feature i to the class c, and $\gamma'_j(f)$ is the FRDD of selected feature j to the new under consideration candidate feature f. The numerator is summation of the FRDD of already selected $k-1$ features as well as newly selected k's feature to the outcome, while the summation in denominator is aggregation of the FRDD of all features except currently under consideration one k's, to itself. It is worth to mention that k in denominator controls the number of selected features. We call the feature selection method based on our proposed merit, Delta Feature Selection (DFS). The numerator can vary from zero to one for each k (since $\gamma'_i \in [0, 1]$), so we have interval of $[0, k]$ in the numerator. However, summation in the denominator varies from zero to $k-1$ for each k, and the whole portion changes in interval of $[\sqrt{k}, k]$ since k is always positive.

The search algorithm of our proposed, that is a greedy forward search method shown in Algorithm 2. The QuickReduct algorithm starts from an empty subset and each time selects one feature to be added to the subset, if the selected feature causes the highest increase in δ; therefore, it will be added to the subset, otherwise, the algorithm evaluates next feature. This process will be continued until no more feature can improve the δ.

To evaluate the applicability of the proposed merit to different types of datasets, a series of criteria have been considered as follows [19]:

Algorithm 2. Delta QuickReduct (DQR)

```
/* $S_f$: best subset of features
   $\delta'_{curr}$: current DFS
   $\delta'_{prev}$: previous DFS
   $nF$: number of features
   $bF$: best feature
*/
$S_f = \{\}$;
$\delta_{curr}, \delta_{prev} = 0$;
do
{
  $\delta_{prev} = \delta_{curr}$;
  for $i = 1$ to $i \leq nF$
  {
    if $\Big((f_i \notin S_f)$ AND $(\delta_{S_f \cup \{f_i\}} > \delta_{prev})\Big)$
    {
      $\delta_{curr} = \delta_{S_f \cup \{f_i\}}$;
      $bF = f_i$;
    }
  }
  $S_f = S_f \cup bF$;
} while $(\delta_{curr} ! = \delta_{prev})$
return $S_f$;
```

1. Correlated and redundant features
2. Non-linearity of data
3. Noisy input
4. Noisy target
5. Small ratio of samples/features
6. Complex datasets

Based on each criteria, thirteen datasets have been adopted from different papers as mentioned in [19] to examine the appropriateness of DFS. Datasts are shown in Table 1. The last column depicts corresponding criterion to the current dataset.

CorrAL dataset has six features, and features one to four are relevant and they generate the outcome by calculating $(f_1 \wedge f_2) \vee (f_3 \wedge f_4)$, feature five is irrelevant, and feature six has 75 % of correlation to the outcome. CorrAL-100 has 99 features that the first six are exactly the same as CorrAL, and the rest are irrelevant and randomly assigned. For both datasets, DFS was able to uncover all four relevant features and also the correlated one.

XOR-100 dataset is a non-linear dataset with two relevant features that compute the output by calculating $(f_1 \oplus f_2)$, and the other 97 features are irrelevant. Again, DFS was able to find two relevant features.

Led-25 dataset is composed of seven relevant features and 17 irrelevant ones. Each dataset, contains the amount of noise (i.e. replacing zero with one or vice

Table 1. Sample datasets to probe different capabilities of a feature selection method

Dataset	#Relevant	#Irrelevant	#Correlated	Criteria
CorrAL [20]	4	1	6	1
CorrAL-100 [21]	4	94	1	1
XOR-100 [21]	2	97	–	2
Led-25 [22] (2 %)	7	17	–	3
Led-25 [22] (6 %)	7	17	–	3
Led-25 [22] (10 %)	7	17	–	3
Led-25 [22] (15 %)	7	17	–	3
Led-25 [22] (20 %)	7	17	–	3
Monk3 [23]	3	3	–	4
SD1 [24]	FCR = 20	4000	–	5
SD2 [24]	FCR = 10, PCR = 30	4000	–	5
SD3 [24]	PCR = 60	4000	–	5
Madelon	5	480	15	6

versa) that is mentioned in parenthesis in front of dataset. Based on the resulting subsets containing two relevant features for all cases, of applying DFS it can be understood that DFS cannot perform well for datasets with noisy inputs.

Monk3 dataset has 5 % of misclassification values as a dataset with noisy target. The DFS has selected features one and five that are irrelevant and relevant, respectively. Therefore, DFS was not able to find all relevant features and also has been misled by noisy target.

SD1, SD2 and SD3 datasets each has three classes, and 75 samples, containing both relevant and irrelevant features. Relevant ones are generated based on a normal distribution, and irrelevant features have been generated based on two distributions namely, normal distribution with mean zero and variance one, and uniform distribution in interval of $[-1, 1]$, each 2000 features. All cancer types can be distinguished by using some genes (or features) called full class relevant (FCR). However, the other genes that are helpful in contrasting some portion of cancer types are called partial class relevant (PCR). Table 2 shows the optimal subset for each dataset, in which nine features out of 10 are redundant features.

Table 2. Optimal features and subsets of SD1, SD2, and SD3

Dataset	#Optimal features/subset	Optimal subsets
SD1 [24]	2	{1–10} {11–20}
SD2 [24]	4	{1–10} {11–20} {21–30} {31–40}
SD3 [24]	6	{1–10} {11–20} {21–30}
		{31–40} {41–50} {51–60}

The DFS has selected 2, 11, and 2 features for SD1, SD2, and SD3, respectively. For SD1, the DFS has selected one feature from the second optimal subset, and one feature from 4000 irrelevant features. For SD2, 11 features have been selected, in which, 10 of them are from the second optimal subset and one feature from 4000 irrelevant features. Finally, two features have been selected from SD3 that one of them is from the third optimal subset and the other one is from irrelevant features.

Madelon dataset has five relevant, 15 linearly correlated to relevant features, and 480 distractor features that are noisy, flipped and shifted [19]. The DFS was able to find five features, in which none of them were among relevant features.

Based on the resulting subsets, our proposed method is capable of dealing with datasets having characteristics mentioned in Table 3.

Table 3. DFS capabilities

Dataset	DFS capability
Correlated and redundant features	✓
Nonlinearity of data	✓
Noisy input	depends on data
Noisy target	depends on data
Small ratio of samples/features	✓
Complex datasets	×

For datasets with noisy input and target, the DFS was capable of finding a subset of relevant features; however, for complex datasets such as Madelon, finding relevant features is very challenging for DFS and many state-of-art feature selection methods [19].

3.2 Performance Measures

In order to evaluate the applicability and performance of FS methods, we define three *Performance* measures to underline classification accuracy and/or reducibility power. The *Reduction* ratio is the value of reduction of total number of features resulting from applying a feature selection method to a datasets, and it is shown in Eq. 13.

$$Reduction = \frac{all\_F - sel\_F}{all\_F}, \tag{13}$$

where all_F is the number of all features, and sel_F is the number of selected features using a feature selection algorithm.

The *Performance* measure is a metric to evaluate the effectiveness of a feature selection algorithm in selecting the smallest subset of features as well as improving the classification accuracy, and is shown by Eq. 14.

$$Performance = CA \times Reduction, \tag{14}$$

where CA is the classification accuracy.

Since the primary aim of FS is to select the smallest meaningful subset of features, we propose a revision of *Performance* measure that emphasizes on the *Reduction* capability of each method and it is presented by Eq. 15.

$$Performance' = CA \times e^{Reduction}, \tag{15}$$

In some cases, data owners prefer those FS methods that lead to higher accuracies; therefore, another revision of Eq. 14 with the aforementioned preference is depicted by Eq. 16.

$$Performance'' = e^{CA} \times Reduction. \tag{16}$$

4 Experimental Results

To validate the proposed method, we have conducted a number of experiments in two steps over 25 UCI [25] traditional as well as newly introduced datasets from three different size categories; Small (S), Medium (M) and Large (L) sizes, in which the number of selected features, *Reduction* ratio, classification accuracy and *Performance* measures are compared. The small size category contains datasets with model size, i.e. $|Features| \times |Samples|$, less than 5 000, the medium size category contains 5 000 to 50 000 cells, and each dataset in the large size category has more than 50 000 cells.

In our experiments the L-FRFS, CFS, and DFS use the same search method called greedy froward search algorithm, and the GBFS uses gradient decent search method.

Computational facilities are provided by ACENET, the regional high performance computing consortium for universities in Atlantic Canada. ACEnet is funded by the Canada Foundation for Innovation (CFI), the Atlantic Canada Opportunities Agency (ACOA), and the provinces of Newfoundland and Labrador, Nova Scotia, and New Brunswick.

4.1 Step One

In this step, we consider all the 25 datasets in our experiment. Table 4 shows the number of samples, features and the size category that each dataset belongs to, and it is sorted based on the model size.

Based on the number of selected features and Eq. 13, the *Reduction* ratio of each method has been calculated and illustrated in Table 5. The cells with zero indicate that the feature selection method could not remove any feature; therefore, all of the features remain untouched.

The bold, superscripted numbers specify the best method in improving the *Reduction* ratio. L-FRFS and GBFS reaches the highest reduction ratio for four datasets, CFS for five datasets, and DFS outperforms the others by gaining the highest *Reduction* values for sixteen datasets. Based on the categories and number of successes of each method, L-FRFS and GBFS result almost similar on the small size category with two and one out of 12 datasets, respectively. However,

Table 4. Datasets specifications

Dataset	Sample	Feature	Size
BLOGGER	100	5	S
Breast Tissue	122	9	S
Qualitative Bankr	250	6	S
Soybean	47	35	S
Glass	214	9	S
Wine	178	13	S
MONK1	124	6	S
MONK2	169	6	S
MONK3	122	6	S
Olitus	120	26	S
Heart	270	13	S
Cleveland	297	13	S
Pima Indian Diab	768	8	M
Breast Cancer	699	9	M
Thoracic Surgery [26]	470	17	M
Climate Model [27]	540	18	M
Ionosphere	351	33	M
Sonar	208	60	M
Wine Quality (Red) [28]	1599	11	M
LSVT Voice Rehab. [29]	126	310	M
Seismic Bumps [30]	2584	18	M
Arrhythmia	452	279	L
Molecular Biology	3190	60	L
COIL 2000 [31]	5822	85	L
Madelon	2000	500	L

DFS highly achieves the best results in both medium and large datasets, by having six out of nine best reduction ratios in medium size category compare to two out of nine for L-FRFS and GBFS methods, and one out of all for CFS. For large datasets, DFS gains 100 % domination. Table 6, shows the number of wins of each method in three categories.

Arithmetic mean has some disadvantages, such as high sensitivity to outliers and also inappropriateness in measuring central tendency of skewed distribution [32], we have conducted the Friedman test that is a non-parametric statistical analysis [33] on the results of Tables 8, 11, 14, and 17 to make the comparison fare enough.

The nine classifiers are PART, Jrip, Naïve Bayes, Bayes Net, J48, BFTree, FT, NBTree, and RBFNetwork that have been selected from different classifier

Table 5. *Reduction* ratio of L-FRFS, CFS, DFS & GBFS

Datasets	L-FRFS	CFS	GBFS	DFS	Size
BLOGGER	0.000	0.400	0.400	**0.600**$^{+}$	S
Breast Tissue	0.000	0.333	**0.444**$^{+}$	0.111	S
Qualitative Bankr.	0.500	0.333	0.167	**0.667**$^{+}$	S
Soybean	**0.943**$^{+}$	0.743	0.886	**0.943**$^{+}$	S
Glass	0.000	0.111	0.333	**0.556**$^{+}$	S
Wine	0.615	0.154	0.692	**0.846**$^{+}$	S
Monk1	0.500	**0.833**$^{+}$	0.333	0.667	S
Monk2	0.000	**0.833**$^{+}$	0.167	0.667	S
Monk3	0.500	**0.833**$^{+}$	0.333	0.667	S
Olitus	**0.808**$^{+}$	0.346	0.731	0.231	S
Heart	0.462	0.462	0.538	**0.846**$^{+}$	S
Cleveland	0.154	**0.923**$^{+}$	0.538	0.846	S
Pima Indian Diab.	0.000	**0.500**$^{+}$	0.250	**0.500**$^{+}$	M
Breast Cancer	0.222	0.000	**0.444**$^{+}$	**0.444**$^{+}$	M
Thoracic Surgery	0.176	0.706	0.588	**0.882**$^{+}$	M
ClimateModel	0.667	0.833	**0.944**$^{+}$	0.889	M
Ionosphere	0.788	0.576	0.818	**0.909**$^{+}$	M
Sonar	**0.917**$^{+}$	0.683	0.900	0.050	M
Wine Quality (Red)	0.000	0.636	0.636	**0.727**$^{+}$	M
LSVT Voice Rehab.	**0.984**$^{+}$	0.900	0.977	0.923	M
Seismic Bumps	0.278	0.667	0.778	**0.889**$^{+}$	M
Arrhythmia	0.975	0.910	0.907	**0.993**$^{+}$	L
Molecular Biology	0.000	0.617	0.000	**0.950**$^{+}$	L
COIL 2000	0.659	0.882	0.941	**0.965**$^{+}$	L
Madelon	0.986	0.982	**0.990**$^{+}$	**0.990**$^{+}$	L

Table 6. Number of wins in achieving the lowest *Reduction* ratio for L-FRFS, CFS, GBFS, and DFS in each category

Algorithm/Category	Small	Medium	Large	Overall
L-FRFS	2	2	0	4
CFS	4	1	0	5
GBFS	1	2	1	4
DFS	6	6	4	16

categories to evaluate the performance of each method by applying 10-fold cross validation (10CV). These classifiers have been implemented in Weka [34], and mean of resulting classification accuracies of all selected classifiers have been used through out the paper. By considering selected features for each dataset, the resulting average of classification accuracies have been shown in Table 7.

By referring to the results in Tables 5 and 7, and applying Eq. 14, the *Performance* measure of each method has been computed and shown in Table 8. The cells that contain zero are the ones with *Reduction* ratio equal to zero. Based on the results shown in Tables 5 and 8, DFS outperforms the other methods by having the best results for 10 datasets compared to that by GBFS with seven, CFS

Table 7. Mean of classification accuracies in % resulting from PART, Jrip, Naïve Bayes, Bayes Net, J48, BFTree, FT, NBTree, and RBFNetwork based on L-FRFS, CFS, GBFS, DFS performance comparing with unreduced datasets

Datasets	L-FRFS	CFS	GBFS	DFS	Unre.
BLOGGER	**74.22**$^+$	73.78	73.78	73.56	**74.22**$^+$
Breast Tissue	**66.46**$^+$	66.35	64.88	65.72	**66.46**$^+$
Qualitative Bankr.	98.44	98.04	98.31	98.40	**98.49**$^+$
Soybean	**100.00**$^+$	75.48	97.87	95.98	98.58
Glass	**67.29**$^+$	66.93	65.42	59.71	61.89
Wine	**95.63**$^+$	95.44	94.63	74.22	85.52
Monk1	**83.13**$^+$	74.07	81.94	73.53	78.32
Monk2	-	67.13	71.89	67.13	**76.62**$^+$
Monk3	98.15	76.23	**98.28**$^+$	75.62	97.92
Olitus	66.39	**75.65**$^+$	53.8	72.69	69.81
Heart	78.48	**81.48**$^+$	81.4	71.32	79.55
Cleveland	49.76	**54.88**$^+$	52.19	54.55	50.13
Pima Indian Diab.	75.00	**75.20**$^+$	**75.20**$^+$	**75.20**$^+$	75.00
Breast Cancer	**96.23**$^+$	96.18	96.23	95.31	96.18
Thoracic Surgery	83.03	84.54	83.95	**85.11**$^+$	83.10
Climate Model	93.25	90.74	91.38	91.36	**93.54**$^+$
Ionosphere	**91.39**$^+$	90.85	89.97	84.96	89.68
Sonar	69.82	**75.48**$^+$	74.89	74.36	67.47
Wine Quality (Red)	58.59	**59.22**$^+$	58.59	56.54	58.59
LSVT Voice Rehab.	**80.60**$^+$	79.37	75.84	72.57	74.69
Seismic Bumps	91.16	91.96	**92.59**$^+$	51.87	91.13
Arrhythmia	53.74	**70.48**$^+$	63.20	59.41	65.46
Molecular Biology	-	73.66	-	51.69	**94.58**$^+$
COIL 2000	92.79	93.65	93.97	**94.02**$^+$	90.61
Madelon	65.79	69.57	**71.27**$^+$	55.26	66.32

with six, and L-FRFS with only three cases. The best performance for small sized datasets has been achieved by DFS and CFS, for medium datasets by DFS and GBFS and for large datasets by DFS. Table 9 evaluates the results of Table 8, and Friedman statistic (distributed according to chi-square with 3 degrees of freedom) is 11.772, and p-value computed by Friedman Test is 0.008206. Based on the rankings, the DFS has gained the best ranking among others; however, its distinction has been examined by post-hoc experiment. The post-hoc procedure as depicted in Table 10 rejects those hypotheses with p-value ≤ 0.030983. So, as shown, DFS and GBFS perform nearly identical. Since performances of DFS and GBFS are not statistically significant, the one with the lowest reduction ratio is selected [35]. Here, based on Table 6, the DFS is ranked the best method among others.

4.2 Step Two

Since the CFS has chosen only one feature for MONK1, MONK2, MONK3 and Cleveland, and also GBFS has selected one out of 18 of Climate Model as the most important feature, further investigations is vital on these suspicious results. The Cleveland dataset has 75 features whereas 13 features out of 75 have been suggested to be used by the published experiments [36]; therefore, all of these 13 features are important from the clinical perspective. By referring to the result of CFS, feature "sex" has been selected as the only important feature due to its highest correlation with the outcome. Neither experts in medical science nor in computer science would arrive at the point that one feature (regardless of type of the feature) out of 13 is enough to predict the outcome. Although selecting "sex" results in the highest classification accuracy, the interpretability of selecting one feature is questionable. Therefore, although "sex" might be an important factor in predicting heart diseases, it is not the only one. For MONK1, MONK2, MONK3 and Climate Model datasets, the only characteristic of the selected feature is its high correlation with the outcome, and very low correlation with the other features.

By removing Cleveland, MONK1, MONK2, MONK3 and Climate Model from Table 8, we form Table 11 and Fig. 1 in which DFS gains the best performance. The GBFS works slightly better than the L-FRFS and CFS for medium datasets, but identical in small datasets. While DFS performance surpasses the GBFS, CFS, and L-FRFS for all three categories. The overall effectiveness and capability of DFS is supported by both Table 11, and the statistical analysis in Table 12. The Friedman statistic (distributed according to chi-square with 3 degrees of freedom) is 9.345, and the p-value computed by Friedman Test is 0.025039. The Li's procedure rejects those hypotheses with p-value ≤ 0.01266, and the results are shown in Table 13. The *Performance* measures resulting form Eqs. 15 and 16 are shown in Tables 14 and 17 and also in Figs. 2 and 3, respectively. The Friedman test results are shown in Tables 15 and 18. For *Performance*$'$, those hypotheses with p-value ≤ 0.00257 are rejected based on Li's procedure, and the results are depicted in Table 16. For *Performance*$''$ as Table 19

Table 8. *Performance* measure resulting from Classification Accuracy × *Reduction*

Datasets	L-FRFS	CFS	GBFS	DFS
BLOGGER	0.000	0.295	0.295	**0.441**$^+$
Breast Tissue	0.000	0.221	**0.288**$^+$	0.073
Qualitative Bankr.	0.492	0.327	0.164	**0.656**$^+$
Soybean	**0.943**$^+$	0.561	0.867	0.905
Glass	0.000	0.074	0.218	**0.332**$^+$
Wine	0.588	0.147	**0.655**$^+$	0.628
Monk1	0.416	**0.617**$^+$	0.273	0.490
Monk2	0.000	**0.559**$^+$	0.120	0.448
Monk3	0.491	**0.635**$^+$	0.328	0.504
Olitus	**0.536**$^+$	0.262	0.393	0.168
Heart	0.362	0.376	0.438	**0.603**$^+$
Cleveland	0.077	**0.507**$^+$	0.281	0.462
Pima Indian Diab.	0.000	**0.376**$^+$	0.188	**0.376**$^+$
Breast Cancer	0.214	0.000	**0.428**$^+$	0.424
Thoracic Surgery	0.147	0.597	0.494	**0.751**$^+$
ClimateModel	0.622	0.756	**0.863**$^+$	0.812
Ionosphere	0.720	0.523	0.736	**0.772**$^+$
Sonar	0.640	0.516	**0.674**$^+$	0.037
Wine Quality (Red)	0.000	0.377	0.373	**0.411**$^+$
LSVT Voice Rehab.	**0.793**$^+$	0.714	0.741	0.670
Seismic Bumps	0.253	0.613	**0.720**$^+$	0.461
Arrhythmia	0.524	**0.642**$^+$	0.573	0.590
Molecular Biology	0.000	0.454	0.000	**0.491**$^+$
COIL 2000	0.611	0.826	0.884	**0.907**$^+$
Madelon	0.649	0.683	**0.706**$^+$	0.547

Table 9. Average rankings of the algorithms based on the *Performance* measure over all datasets (Friedman)

Algorithm	Ranking
L-FRFS	3.220
CFS	2.440
GBFS	2.320
DFS	**2.020**$^+$

Table 10. Post Hoc comparison over the results of Friedman procedure of *Performance* measure

i	Algorithm	$z = (R_0 - R_i)/SE$	p	Li
3	L-FRFS	3.286335	0.001015	0.030983
2	CFS	1.150217	0.250054	0.030983
1	GBFS	0.821584	0.411314	0.05

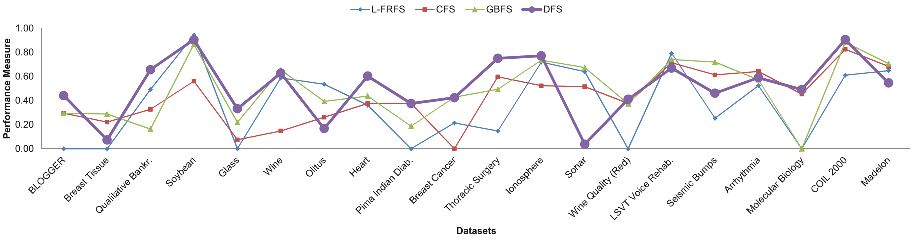

Fig. 1. *Performance* measure (Classification Accuracy × *Reduction*)

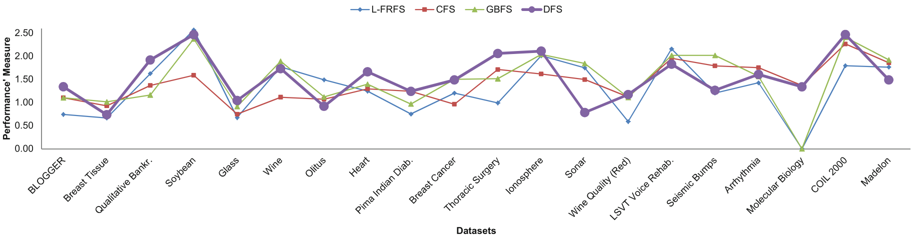

Fig. 2. *Performance′* measure (Classification Accuracy $\times e^{Reduction}$)

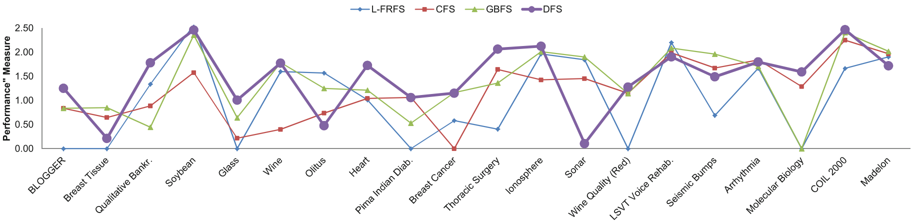

Fig. 3. *Performance″* measure ($e^{ClassificationAccuracy} \times$ *Reduction*)

shows, those hypotheses with p-value ≤ 0.01266 are rejected based on Li's procedure. Figures 1, 2 and 3 depict *Performance*, *Performance′* and *Performance″* measures values for each dataset, respectively.

Table 11. *Performance* measure resulting from classification accuracy × *reduction* after removing Cleveland, MONK1, MONK2, MONK3 & Climate Model

Datasets	L-FRFS	CFS	GBFS	DFS
BLOGGER	0.000	0.295	0.295	**0.441**$^+$
Breast Tissue	0.000	0.221	**0.288**$^+$	0.073
Qualitative Bankr.	0.492	0.327	0.164	**0.656**$^+$
Soybean	**0.943**$^+$	0.561	0.867	0.905
Glass	0.000	0.074	0.218	**0.332**$^+$
Wine	0.588	0.147	**0.655**$^+$	0.628
Olitus	**0.536**$^+$	0.262	0.393	0.168
Heart	0.362	0.376	0.438	**0.603**$^+$
Pima Indian Diab.	0.000	**0.376**$^+$	0.188	**0.376**$^+$
Breast Cancer	0.214	0.000	**0.428**$^+$	0.424
Thoracic Surgery	0.147	0.597	0.494	**0.751**$^+$
Ionosphere	0.720	0.523	0.736	**0.772**$^+$
Sonar	0.640	0.516	**0.674**$^+$	0.037
Wine Quality (Red)	0.000	0.377	0.373	**0.411**$^+$
LSVT Voice Rehab.	**0.793**$^+$	0.714	0.741	0.670
Seismic Bumps	0.253	0.613	**0.720**$^+$	0.461
Arrhythmia	0.524	**0.642**$^+$	0.573	0.590
Molecular Biology	0.000	0.454	0.000	**0.491**$^+$
COIL 2000	0.611	0.826	0.884	**0.907**$^+$
Madelon	0.649	0.683	**0.706**$^+$	0.547

Table 12. Average rankings of the algorithms based on the *Performance* measure after removing Cleveland, MONK1, MONK2, MONK3 & Climate Model (Friedman)

Algorithm	Ranking
L-FRFS	3.125
CFS	2.700
GBFS	2.150
DFS	**2.025**$^+$

Table 13. Post Hoc comparison over the results of Friedman procedure of *Performance* measure after removing Cleveland, MONK1, MONK2, MONK3 & Climate Model

i	Algorithm	$z = (R_0 - R_i)/SE$	p	Li
3	L-FRFS	2.694439	0.007051	0.01266
2	CFS	1.653406	0.098248	0.01266
1	GBFS	0.306186	0.759463	0.05

Table 14. *Performance′* measure resulting from Classification Accuracy $\times e^{Reduction}$

Datasets	L-FRFS	CFS	GBFS	DFS
BLOGGER	0.742	1.101	1.101	**1.340**$^+$
Breast Tissue	0.665	0.926	**1.012**$^+$	0.734
Qualitative Bankr.	1.623	1.368	1.161	**1.917**$^+$
Soybean	**2.567**$^+$	1.587	2.373	2.464
Glass	0.673	0.748	0.913	**1.041**$^+$
Wine	1.770	1.113	**1.891**$^+$	1.730
Olitus	**1.489**$^+$	1.069	1.117	0.916
Heart	1.245	1.293	1.395	**1.662**$^+$
Pima Indian Diab.	0.750	**1.240**$^+$	0.966	**1.240**$^+$
Breast Cancer	1.202	0.962	**1.501**$^+$	1.487
Thoracic Surgery	0.990	1.712	1.512	**2.057**$^+$
Ionosphere	2.009	1.616	2.039	**2.109**$^+$
Sonar	1.746	1.495	**1.842**$^+$	0.782
Wine Quality (Red)	0.586	1.119	1.107	**1.170**$^+$
LSVT Voice Rehab.	**2.156**$^+$	1.952	2.015	1.826
Seismic Bumps	1.204	1.791	**2.015**$^+$	1.262
Arrhythmia	1.425	**1.752**$^+$	1.565	1.604
Molecular Biology	0.000	**1.365**$^+$	0.000	1.337
COIL 2000	1.793	2.263	2.408	**2.467**$^+$
Madelon	1.763	1.857	**1.918**$^+$	1.487

Table 15. Average rankings of the algorithms based on the *Performance′* measure after removing Cleveland, MONK1, MONK2, MONK3 & Climate Model (Friedman)

Algorithm	Ranking
L-FRFS	3.075
CFS	2.650
DFS	2.150
GBFS	**2.125**$^+$

Table 16. Post Hoc comparison over the results of Friedman procedure of *Performance′* measure after removing Cleveland, MONK1, MONK2, MONK3 & Climate Model

i	Algorithm	$z = (R_0 - R_i)/SE$	p	Li
3	L-FRFS	2.327015	0.019964	0.00257
2	CFS	1.285982	0.198449	0.00257
1	GBFS	0.061237	0.95117	0.05

Table 17. *Performance''* measure resulting from $e^{ClassificationAccuracy} \times$ *Reduction*

Datasets	L-FRFS	CFS	GBFS	DFS
BLOGGER	0.000	0.837	0.837	**1.252**$^{+}$
Breast Tissue	0.000	0.647	**0.850**$^{+}$	0.214
Qualitative Bankr.	1.338	0.889	0.445	**1.783**$^{+}$
Soybean	**2.563**$^{+}$	1.580	2.357	2.462
Glass	0.000	0.217	0.641	**1.009**$^{+}$
Wine	1.601	0.400	**1.784**$^{+}$	1.777
Olitus	**1.569**$^{+}$	0.738	1.251	0.477
Heart	1.012	1.043	1.215	**1.727**$^{+}$
Pima Indian Diab.	0.000	**1.061**$^{+}$	0.530	**1.061**$^{+}$
Breast Cancer	0.582	0.000	**1.163**$^{+}$	1.153
Thoracic Surgery	0.405	1.644	1.362	**2.067**$^{+}$
Ionosphere	1.965	1.428	2.012	**2.126**$^{+}$
Sonar	1.843	1.454	**1.903**$^{+}$	0.105
Wine Quality (Red)	0.000	1.151	1.143	**1.280**$^{+}$
LSVT Voice Rehab.	**2.203**$^{+}$	1.990	2.087	1.906
Seismic Bumps	0.691	1.672	**1.963**$^{+}$	1.493
Arrhythmia	1.669	**1.842**$^{+}$	1.706	1.799
Molecular Biology	0.000	1.288	0.000	**1.593**$^{+}$
COIL 2000	1.666	2.251	2.409	**2.470**$^{+}$
Madelon	1.904	1.969	**2.019**$^{+}$	1.720

Table 18. Average rankings of the algorithms based on the *Performance''* measure after removing Cleveland, MONK1, MONK2, MONK3 & Climate Model (Friedman)

Algorithm	Ranking
L-FRFS	3.125
CFS	2.700
GBFS	2.150
DFS	**2.025**$^{+}$

Table 19. Post Hoc comparison over the results of Friedman procedure of *Performance''* measure after removing Cleveland, MONK1, MONK2, MONK3 & Climate Model

i	Algorithm	$z = (R_0 - R_i)/SE$	p	Li
3	L-FRFS	2.694439	0.007051	0.01266
2	CFS	1.653406	0.098248	0.01266
1	GBFS	0.306186	0.759463	0.05

5 Conclusions and Future Work

This paper introduces a new hybrid merit based on conjunction of correlation feature selection and fuzzy-rough feature selection. It takes advantages of both methods by integrating them into a new hybrid merit to improve the quality of the selected subsets as well as resulting reasonable classification accuracies. The new merit selects less number of redundant features, and finds the most relevant ones to the outcome.

The performance of the proposed merit is examined with a variety of different datasets with diverse number of features and samples, that have been chosen because of their predominance as well as recently introduced in the literature. The two-step experimental results show the effectiveness of our new hybrid merit over divergent UCI datasets, especially on medium and large ones. We have also proposed three measures to thoroughly figure out and compare the performance of feature selection methods.

Based on the results, we conclude that proposing a universal feature selection method might not be suitable due to the high variety of datasets and applications. Therefore, each and every newly proposed method can be "localized" to a subject and type of the data as well as the purpose of the data. In such a way, data owners can save huge amounts of processing expenses based on a set of categorized methods. As future work, we are excited to perform such categorization for the existing merits on feature selection methods. Also, we are conducting some experiments on Big Data in order to evaluate the performance of the proposed hybrid merit.

Our ongoing task is to prepare an online, web-based application for the new hybrid merit that will be available to the researchers working on datasets in various field of studies.

Acknowledgments. This work has been partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Research & Development Corporation of Newfoundland and Labrador (RDC).

References

1. Hall, M.A., Smith, L.A.: Feature subset selection: a correlation based filter approach. In: Proceedings of the 1997 International Conference on Neural Information Processing and Intelligent Information Systems, New Zealand, pp. 855–858 (1997)
2. Javed, K., Babri, H.A., Saeed, M.: Feature selection based on class-dependent densities for high-dimensional binary data. IEEE Trans. Knowl. Data Eng. **24**, 465–477 (2012)
3. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artif. Intell. **97**, 273–324 (1997)
4. Das, S.: Filters, wrappers and a boosting-based hybrid for feature selection. In: ICML, vol. 1, pp. 74–81. Citeseer (2001)
5. Kira, K., Rendell, L.A.: The feature selection problem: traditional methods and a new algorithm. In: AAAI, pp. 129–134 (1992)

6. Jensen, R., Shen, Q.: New approaches to fuzzy-rough feature selection. IEEE Trans. Fuzzy Syst. **17**, 824–838 (2009)
7. Anaraki, J.R., Eftekhari, M., Ahn, C.W.: Novel improvements on the fuzzy-rough quickreduct algorithm. IEICE Trans. Inf. Syst. **E98.D**(2), 453–456 (2015)
8. Anaraki, J.R., Eftekhari, M.: Improving fuzzy-rough quick reduct for feature selection. In: 2011 19th Iranian Conference on Electrical Engineering (ICEE), pp. 1502–1506 (2011)
9. Qian, Y., Wang, Q., Cheng, H., Liang, J., Dang, C.: Fuzzy-rough feature selection accelerator. Fuzzy Sets Syst. **258**, 61–78 (2015). Special issue: Uncertainty in Learning from Big Data
10. Jensen, R., Vluymans, S., Parthaláin, N.M., Cornelis, C., Saeys, Y.: Semi-supervised fuzzy-rough feature selection. In: Yao, Y., Hu, Q., Yu, H., Grzymala-Busse, J.W. (eds.) RSFDGrC 2015. LNCS (LNAI), vol. 9437, pp. 185–195. Springer, Heidelberg (2015). doi:10.1007/978-3-319-25783-9_17
11. Shang, C., Barnes, D.: Fuzzy-rough feature selection aided support vector machines for mars image classification. Comput. Vis. Image Underst. **117**, 202–213 (2013)
12. Derrac, J., Verbiest, N., García, S., Cornelis, C., Herrera, F.: On the use of evolutionary feature selection for improving fuzzy rough set based prototype selection. Soft Comput. **17**, 223–238 (2012)
13. Dai, J., Xu, Q.: Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification. Appl. Soft Comput. **13**, 211–221 (2013)
14. Xu, Z., Huang, G., Weinberger, K.Q., Zheng, A.X.: Gradient boosted feature selection. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 522–531. ACM (2014)
15. Yu, L., Liu, H.: Efficient feature selection via analysis of relevance and redundancy. J. Mach. Learn. Res. **5**, 1205–1224 (2004)
16. Pawlak, Z.: Rough sets. Int. J. Comput. Inf. Sci. **11**, 341–356 (1982)
17. Komorowski, J., Pawlak, Z., Polkowski, L., Skowron, A.: Rough sets: a tutorial. In: Pal, S.K., Skowron, A. (eds.) Rough-Fuzzy Hybridization: A New Trend in Decision Making, pp. 3–98. Springer-Verlag New York, Inc., Secaucus (1998)
18. Radzikowska, A.M., Kerre, E.E.: A comparative study of fuzzy rough sets. Fuzzy Sets Syst. **126**, 137–155 (2002)
19. Boln-Canedo, V., Snchez-Maroo, N., Alonso-Betanzos, A.: Feature Selection for High-Dimensional Data. Springer, Switzerland (2016)
20. John, G.H., Kohavi, R., Pfleger, K., et al.: Irrelevant features and the subset selection problem. In: Machine Learning: Proceedings of the Eleventh International Conference, pp. 121–129 (1994)
21. Kim, G., Kim, Y., Lim, H., Kim, H.: An mlp-based feature subset selection for HIV-1 protease cleavage site analysis. Artif. Intell. Med. **48**, 83–89 (2010). Artificial Intelligence in Biomedical Engineering and Informatics
22. Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A.: Classification and regression trees. CRC Press, New York (1984)
23. Wnek, J., Michalski, R.S.: Comparing symbolic and subsymbolic learning: three studies. Mach. Learn. A Multistrategy Approach **4**, 318–362 (1994)
24. Zhu, Z., Ong, Y.S., Zurada, J.M.: Identification of full and partial class relevant genes. IEEE/ACM Trans. Comput. Biol. Bioinform. **7**, 263–277 (2010)
25. Bache, K., Lichman, M.: UCI machine learning repository (2013)
26. Zieba, M., Tomczak, J.M., Lubicz, M., Swiatek, J.: Boosted svm for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. Appl. Soft Comput. **14**, 99–108 (2014)

27. Lucas, D.D., Klein, R., Tannahill, J., Ivanova, D., Brandon, S., Domyancic, D., Zhang, Y.: Failure analysis of parameter-induced simulation crashes in climate models. Geoscientific Model Devel. **6**, 1157–1171 (2013)
28. Cortez, P., Cerdeira, A., Almeida, F., Matos, T., Reis, J.: Modeling wine preferences by data mining from physicochemical properties. Decis. Support Syst. **47**, 547–553 (2009)
29. Tsanas, A., Little, M., Fox, C., Ramig, L.: Objective automatic assessment of rehabilitative speech treatment in parkinson's disease. IEEE Trans. Neural Syst. Rehabil. Eng. **22**, 181–190 (2014)
30. Sikora, M., Wróbel, Ł.: Application of rule induction algorithms for analysis of data collected by seismic hazard monitoring systems in coal mines. Arch. Min. Sci. **55**, 91–114 (2010)
31. Putten, P.V.D., Someren, M.V.: Coil challenge 2000: the insurance company case. Technical report 2000–2009. Leiden Institute of Advanced Computer Science, Universiteit van Leiden (2000)
32. Manikandan, S.: Measures of central tendency: the mean. J. Pharmacol. Pharmacotherapeutics **2**, 140 (2011)
33. Alcala-Fdez, J., Fernandez, A., Luengo, J., Derrac, J., Garcia, S.: Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. Multiple-Valued Logic Soft Comput. **17**, 255–287 (2011)
34. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. SIGKDD Explor. Newsl. **11**, 10–18 (2009)
35. Guyon, I., Gunn, S., Ben-Hur, A., Dror, G.: Result analysis of the nips 2003 feature selection challenge. In: Advances in Neural Information Processing Systems, pp. 545–552 (2004)
36. Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.J., Sandhu, S., Guppy, K.H., Lee, S., Froelicher, V.: International application of a new probability algorithm for the diagnosis of coronary artery disease. Am. J. Cardiol. **64**, 304–310 (1989)